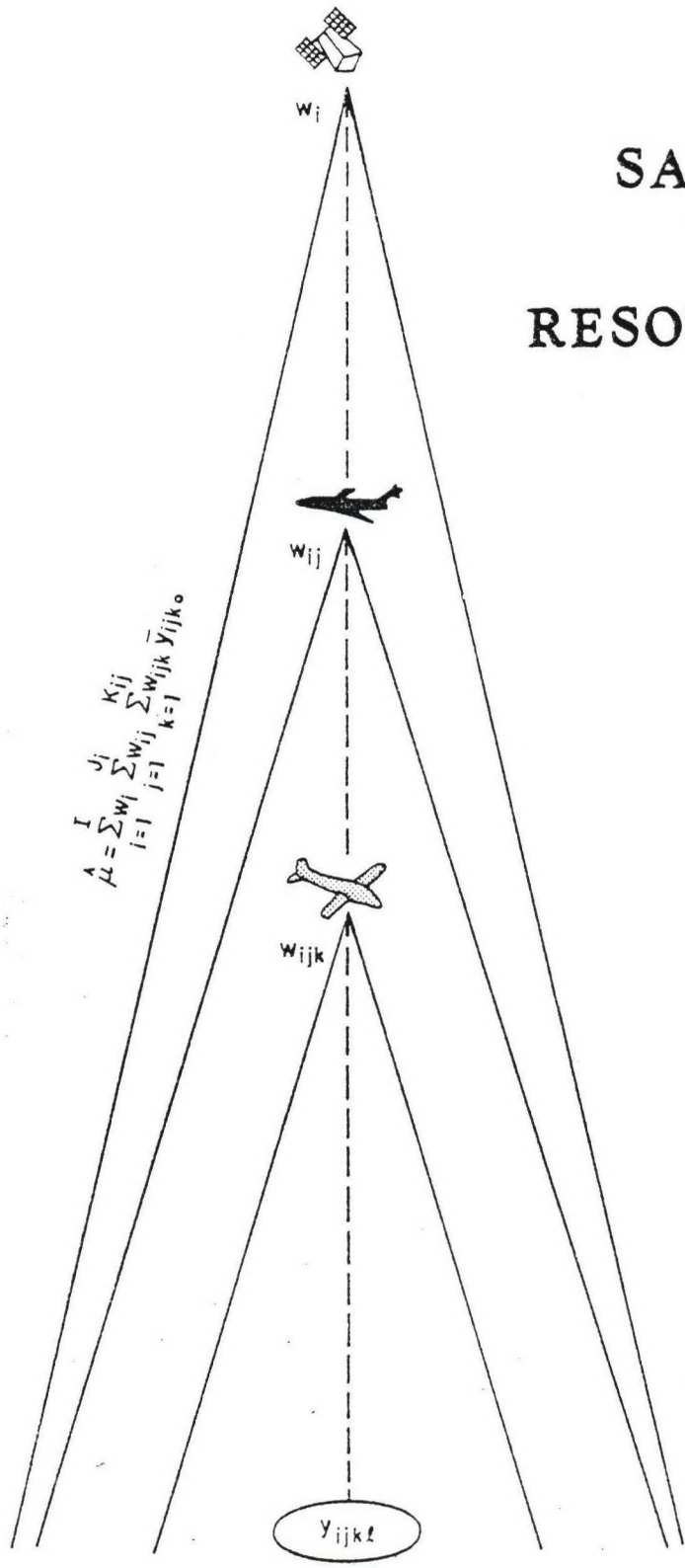
MULTI-LEVEL SAMPLING DESIGNS FOR RESOURCE INVENTORIES

NOVEMBER 1979



MULTI-LEVEL SAMPLING DESIGNS FOR RESOURCE INVENTORIES

November 1979

Department of
Forest and Wood Sciences
Colorado State University

Rocky Mountain
Forest and Range
Experiment Station
USDA Forest Service

ABSTRACT

Multiphase and multistage sampling through four levels was examined. Options involved simple random and systematic subsampling, point and interval estimation, and unbiased and ratio estimators. Formulas were derived for estimating means, totals, and proportions, and the variances associated with these estimates. Formulas for sample size computation and allocation were also derived. Examples using data from Grand County, Colorado are given.

PREFACE

In the summer of 1977, I approached Dick Driscoll of the U.S. Forest Service Rocky Mountain Forest and Range Experiment Station to suggest a cooperative research study on multi-level sampling designs. The study began shortly after that initial discussion. When the study plan was completed, it indicated that multi-stage and multi-phase designs would be examined. A major purpose for choosing these designs was to explore the use of satellite imagery and selected types of aerial photography in a sampling design framework.

Very few applications of such designs had been made previously in natural resource inventories with two exceptions: 1) double sampling using a combination of aerial photographs and ground plots had been applied for many years, and 2) two-stage sampling had been studied extensively by Phil Langley, Earth Satellite Corp., in conjunction with Lee Wensel, Univ. of California - Berkeley.

Major contributors to the study from the Rocky Mountain Station include Gary Dixon, Jim LaBau and Robert Aldrich. Others at the Station were also involved in collecting and compiling data used for examples in the study.

In addition to myself, contributors to the study from Colorado State University include Frank Graybill, Professor of Statistics; Sakthivel Jeyaratnam (known as "Jay" to those involved in the study), a former post-doctoral Fellow in the Department of Statistics; Brian Kent, Assistant Professor of Forest Biometry; Dave Johnston, PhD candidate in the Department of Forest and Wood Sciences; and Dave Bowden, Associate Professor of Statistics, who reviewed the draft of the final report and made many constructive suggestions for improvement.

As with any document this size of a mathematical nature, it is virtually impossible to eliminate all errors. I must assume responsibility for such errors as may exist; accordingly, I shall appreciate being notified of any that you may discover.

W. E. Frayer
November, 1979

Major contributors to the study:

(In alphabetical order)

David C. Bowden

Gary E. Dixon

W. E. Frayer

Franklin A. Graybill

Sakthivel Jeyaratnam

David C. Johnston

Brian M. Kent

Vernon J. LaBau

Ed Roberts

REPORT

RM Contract 16-747-CA

CSU Project 31-1470-1468

September 30, 1979

MULTI-LEVEL SAMPLING DESIGNS FOR RESOURCE INVENTORY

PART I

BACKGROUND

Chapter 1. INTRODUCTION

1.1 Discussion. In natural resource studies, it is often necessary to obtain measurements on a set of items when it is very expensive or virtually impossible to measure the entire set. For example, it may be necessary to know the average or total weight of all cattle on a large range; it may be desirable to know the total amount of merchantable wood in a certain area; it may be useful to know the total biomass produced in a fishery; it may be necessary to know the proportion of dead trees in a forest; etc. In each of these examples it would be extremely expensive (or perhaps impossible) to obtain the measurement of each item under study (the weight of each head of cattle, the volume of each tree, the weight of each fish, the proportion of dead trees in the forest; etc.). One approach is to measure only a few of the items under study and on the basis of these measurements, make an inference about the entire set. If the small subset is selected in a specified manner, the inference may be useful and reliable.

Some of the important problems that arise in studies of these kinds are stated here: 1) How should the subset (which is called the sample) be selected? 2) How many items should be included in the sample? 3) What formulas should be used to approximate (estimate) the quantities in the entire set (population)? 4) How can the error of estimation be evaluated? The subject that includes these types of questions is called "survey sampling" and this monograph is devoted to a specialized treatment of some survey sampling methods that have been found useful in natural resource studies.

1.2 Definitions and Notations. In this section basic definitions and notations will be stated; additional definitions and notations will be given in later chapters as needed.

To begin, suppose a study is undertaken in order to make a quantitative (numerical) statement about a set of measurements. The set of measurements is

the population under investigation and the definition is given below.

> Definition 1.1.1 Universe and Population. A universe is a set of items under study. A population is a set of numbers where each number is the measurement, or indicator number, of a characteristic of each item in the universe.

Note. An illustration of a population of indicator numbers is given in Example 1.2.3.

It is extremely important that the population be rigorously defined. The population size N may be known, but it is often unknown. Some examples of populations are given below.

Example 1.2.1. The population of weights of cattle in a certain large area might be the set of numbers (in pounds)--{600, 824, 993, ..., 815}. From this population, the average or total weight of the animals could be computed.

Example 1.2.2. The population of the amount of merchantable wood on each acre of a large area might be the set of numbers (in cubic feet)--{2541, 3802, 1695, 5947, ..., 8580}. From this population, the total amount of merchantable wood can be obtained.

Example 1.2.3. To determine the proportion of dead trees in a certain forest, the population could be the following set--{1, 1, 0, 1, 1, 1, 0, 0, 1, ..., 1} where each tree is assigned the number 0 if it is not dead, the number 1 if it is dead. This is an example of a population where the numbers are indicator numbers; a 0 indicates a tree is alive, a 1 indicates the tree is dead. From this population, the proportion of dead trees can be obtained.

As we stated, the population is almost never known in a real problem, but it can (and must) be rigorously defined. In general, it is not essential to be able to determine all the numbers in a population, but rather what one usually wants are summary numbers obtained from the population. The most important summary numbers, called population parameters, are the population mean; population

total, population proportion, population variance, and population standard deviation. The numbers in a population are denoted by capital latin letters; i.e. by Y_1, Y_2, etc. So a population consisting of N numbers is denoted by

$$\{Y_1, Y_2, \ldots, Y_N\}.$$

The summary numbers mentioned above are defined in the table below.

Table 1.2.1

Population Parameter	Symbol	Definition
Mean	μ or $\bar{Y}$	$\mu = \frac{1}{N} \sum_{i=1}^{N} Y_i$
Total	τ or Y	$\tau = \sum_{i=1}^{N} Y_i$
Proportion	p	$p = \frac{1}{N} \sum_{i=1}^{N} Y_i$ when Y_i is an indicator number
Variance [1/]	S^2 or σ^2	$S^2 = (N-1)^{-1} \sum_{i=1}^{N} \left(Y_i - \bar{Y}\right)^2$
Standard deviation	S or σ	$S = \sqrt{S^2}$

1.3 Samples. We have stated that in a real-world situation, it is usually impossible or impractical to know the population, but it is also true that what is generally required is a knowledge of certain summary parameters. It is, however, also impossible to be able to calculate the parameters since to do so would require a knowledge of all of the population values (the Y_i for $i = 1, \ldots, N$). Because all of the population values can usually not be known, a few are obtained and from these one estimates the population parameters. These few that are observed are defined to be sample valuesfrom the population and the sample is defined below.

1/ Some authors define the population variance as $S^2 = N^{-1} \sum_{i=1}^{N} (Y_i - \bar{Y})^2$, but the formulas simplify if it is defined as in the Table (Cochran, 1976).

Definition 1.3.1 Sample of Size n. A set of n numbers from a population which is obtained in a prescribed manner is called a sample of size n from the population. The sample values are denoted by lower case latin letters (say y_i) so the sample is denoted by $y_1, y_2, \ldots, y_n$.

The definition states that the sample values must be obtained from the population in a certain prescribed manner. The method of obtaining the sample determines the type of sample that it is, i.e. 1) simple random sample; 2) cluster sample; 3) stratified random sample; etc. The methods used in this monograph will be defined when they are discussed in later chapters. From the known sample values one can estimate the unknown population parameters and "determine" how "good" the estimators are. These concepts will be discussed in the next section.

1.4 Point Estimates. From a population $\{Y_1, Y_2, \ldots, Y_N\}$ suppose it is desired to estimate a parameter denoted by θ where θ could be the population mean μ, the population total τ, or any other parameter. A sample of size n is selected from the population and denoted by $y_1, y_2, \ldots, y_n$. An estimate of θ which is denoted by $\hat{\theta}$ is a known number computed from the sample values by a specified formula, and $\hat{\theta}$ is called a point estimate of the parameter θ. The particular formula used is determined from theoretical statistical procedures so that $\hat{\theta}$ is a "good" estimate of θ where "good" is defined from certain intuitive concepts. In survey sampling, a "good" estimate is generally required to be unbiased and have a small average squared error. Unbiased means that on the average, $\hat{\theta}$ is equal to θ. This means that if a value of the estimate $\hat{\theta}$ is computed for every possible sample of size n that could be obtained from the population, then the "average" of all of these estimates is exactly equal to the parameter θ. This is stated formally in the following definition.

Definition 1.4.1 Unbiased Estimator. An estimator $\hat{\theta}$ of a parameter θ is defined to be an unbiased estimator of θ if the average of all the $\hat{\theta}$'s (computed from all possible samples of size n from the population) is equal to θ. The symbol used for the average of all the values of $\hat{\theta}$ is $\mathcal{E}[\hat{\theta}]$ where $\mathcal{E}$ is a symbol for average. This is an unbiased estimator of θ if:

$$\mathcal{E}[\hat{\theta}] = \theta.$$

Of course, we want the estimate $\hat{\theta}$ to be "close" to the parameter θ, but we know that in general $\hat{\theta}$ will never be exactly equal to θ (even if it is, we would not know this since θ is unknown). The error of estimation of θ using $\hat{\theta}$ is $\hat{\theta}-\theta$, and the square of the error of estimation is $(\hat{\theta}-\theta)^2$. We want this to be small, but it depends on the particular sample values obtained. Consequently, to determine how "good" the estimator $\hat{\theta}$ is we need to consider the magnitude of the "average" square of the error. The smaller "the average square of the error" the better the estimate. The "average squared error" (also called "mean squared error") is: $\mathcal{E}[(\hat{\theta} - \theta)^2]$.

The concept of "error of estimation" is formally defined below.

Definition 1.4.2 Mean Squared Error. If $\hat{\theta}$ is an estimate of an unknown population parameter θ, then $\mathcal{E}[(\hat{\theta}-\theta)^2]$ is defined as the mean squared error of estimating θ; $\hat{\theta}-\theta$ is defined as the error of estimating θ, and $|\hat{\theta}-\theta|$ is called the absolute error of estimating θ.

Note. If $\hat{\theta}$ is an unbiased estimate of θ, then the mean squared error $\mathcal{E}[(\hat{\theta}-\theta)]^2$ is the variance of $\hat{\theta}$.

Note. In this monograph, an estimate $\hat{\theta}$ of a parameter θ that is unbiased and has a variance that is smaller than any other unbiased estimate will be called the best unbiased estimate. However, an estimator may be used that is not unbiased since a biased estimate may have a smaller mean squared error than an unbiased estimate.

1.5 Interval Estimation. A point estimate of an unknown population parameter θ gives very important information. However, a point estimate by itself does not give information about how "close" the sample estimate $\hat{\theta}$ is to the population parameter θ. A confidence interval about θ does give this information in terms of probability. For example, consider a $1-\alpha$ confidence interval on θ:

$$L\left(y_1, y_2, \ldots, y_n\right) \leq \theta \leq U\left(y_1, y_2, \ldots, y_n\right).$$

$L(y_1, y_2, \ldots, y_n)$ and $U(y_1, y_2, \ldots, y_n)$ are respectively the lower and upper ends of the confidence interval and they are known functions of the observed sample values $y_1, y_2, \ldots, y_n$. The meaning of a confidence interval is this: each possible sample of size n from a population of size N would give a confidence interval on θ. The proportion of the totality of all possible confidence intervals that includes the unknown parameter is equal to the specified confidence coefficient $1-\alpha$. Of course in an applied problem, only one confidence interval will be computed but this one can be considered as a random sample of size 1 from all possible confidence intervals that could have been computed (one for each possible sample). Since $1-\alpha$ of all possible confidence intervals include θ, the probability is equal to $1-\alpha$ that the one actually computed contains θ.

For many problems that will be considered in this monograph, an approximate $1-\alpha$ confidence interval on θ is of the form:

$$\hat{\theta} - t_\gamma s_{\hat{\theta}} \leq \theta \leq \hat{\theta} + t_\gamma s_{\hat{\theta}}$$

where t_γ is an appropriate tabled value (such as a value from a table of normal deviates, a value from Student's t table, etc.), γ depends on the specified confidence coefficient $1-\alpha$ (γ is generally equal to $\alpha/2$), and $s_{\hat{\theta}}$ is the sample standard deviation of the estimate $\hat{\theta}$. The confidence coefficient $1-\alpha$ is generally specified to be one of the values .80, .90, .95, .99. The formulas for computing $\hat{\theta}$ and $s_{\hat{\theta}}$ will be exhibited for each sampling design and each parameter of interest.

In evaluating information provided by a confidence interval, there are two important things to consider--the confidence coefficient and the width (or expected width) of the interval. The confidence coefficient determines the probability that the interval actually contains the unknown parameter θ, and the width tells how far the estimator $\hat{\theta}$ is from the parameter θ. If the width is too large, the confidence coefficient may not be useful; if it is smaller than required to make a decision, then resources have been wasted and a decision could have been made at a smaller cost. In general, for the types of problems to be discussed in this monograph the width depends on the sample size n such that as n increases the width tends to decrease. So it will be useful to determine the sample size that should be used to obtain the confidence interval width that is required. These ideas are illustrated with a specific example.

Example 1.5.1 Suppose 2,000 head of cattle are to be brought from summer range to a feedlot and it is desired to determine the average weight of the cattle. One does not want to weigh all 2,000 of the cattle so a random sample of 16 of the cattle are selected and weighed. The sample mean is $\bar{y} = 625$, and the sample standard deviation is $s = 220$. An approximate $1-\alpha = .95$ confidence interval on the average weight of the 2,000 cattle is:

$$\bar{y}-2.13(s/\sqrt{n}) \leq \mu \leq \bar{y}+2.13(s\sqrt{n}) \quad \text{where } s_{\hat{\theta}} = s/\sqrt{n}$$

which is $$625-2.13(220/4) \leq \mu \leq 625+2.13(220/4)$$

$$625-117 \leq \mu \leq 625+117$$

$$508 \leq \mu \leq 742$$

The width of the confidence interval is 742-508 = 234. If the interval is too wide to be useful, then one can take additional sample values so that the expected width will be approximately a value specified by the investigator. The

formulas for the appropriate values of n, the sample size, will be given for each design as it is discussed.

1.6 Two-Variable Population. In some investigations, the population consists of a set of ordered pairs of numbers since there are often two characteristics of importance in the universe under study. This idea is introduced with an example.

Example 1.6.1 In Example 1.5.1 suppose the two variables under study for each of the 2,000 cattle are 1) weight when they were put on summer range, X; 2) weight after they were brought from summer range, Y. So the population consists of 2,000 pairs of numbers and could be displayed as:

$$\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_{2000}, Y_{2000})\}.$$

One of the things that the investigator would like to know is R the ratio of total weight at the end of the summer to the total weight at the beginning of the summer defined by:

$$R = \sum_{i=1}^{N} Y_i \Big/ \sum_{i=1}^{N} X_i ,$$

The population values are not known so sample values will be used to estimate R. These ideas can be easily extended to more than two variables.

Chapter 2. SIMPLE RANDOM SAMPLING

2.1 Definition and Notation. In order to estimate the various parameters (mean, total, variance, ratio, etc.) in a population, sample values will be obtained and estimates will be formed as functions of these known numbers.

There are many ways to select the sample from the population, and the investigator should choose a method of selection that will result in satisfactory estimates for a minimum cost. The various methods of selecting the sample are referred to as "sample designs," and several methods that have been found useful in natural resource work will be discussed in this monograph. The formulas used for estimating the parameters of interest will depend on the sample design used to select the sample.

An easy design to understand and one that is often used in conjunction with other, more complicated designs, is called a "simple random sampling" design.

Definition 2.1.1 Simple Random Sample. If a universe consists of N items and if a sample of size n is selected from this universe so that every distinct sample of size n has the same chance of being selected, the sample is defined to be a simple random sample and the design is defined to be a simple random sampling design.

There are many ways of obtaining a sample that guarantees that it is a simple random sample. If all of the items in a universe can be easily identified and numbered from 1 to N, then a table of random numbers can be used (see Cochran Secs. 2.1 and 2.2 for details). The n sample values will be denoted by $y_1, y_2, \ldots, y_n$.

2.2 Point and Interval Estimates. In this section, the formulas for estimating population parameters are given when the sample is obtained by simple random sampling. Table 2.2.1 contains a summary of the point estimates of the population mean μ, total τ, and proportion p.

Table 2.2.1

Population Parameter	Point Estimate	Variance of Estimate	Estimate of Variance
μ	$\hat{\mu}=\bar{y}$	$V(\hat{\mu}) = (S^2/n)(1-f)$	$\hat{V}(\hat{\mu}) = (s^2/n)(1-f)$
τ	$\hat{\tau}=N\bar{y}$	$V(\hat{\tau}) = (N^2S^2/n)(1-f)$	$\hat{V}(\hat{\tau}) = (N^2s^2/n)(1-f)$
p	$\hat{p}=\bar{y}$	$V(\hat{p}) = \frac{pq}{n-1}(1-f)$	$\hat{V}(\hat{p}) = \frac{\hat{p}\hat{q}}{n-1}(1-f)$

where $\bar{y} = (1/n) \sum_{i=1}^{n} y_i$, where $q = 1-p$, where $\hat{q} = 1-\hat{p}$, where the symbol used for the variance of an estimator $\hat{\theta}$ is $V(\hat{\theta})$, and the estimated variance is $\hat{V}(\hat{\theta})$. Also, $f = n/N$, the ratio of sample size to population size, is called the sampling fraction; s^2, the sample variance is defined by:

$$s^2 = (n-1)^{-1} \sum_{i=1}^{n} (y_i-\bar{y})^2.$$

Note The definition of S^2, the population variance, is given in Table 1.2.1.

2.3 Confidence Interval Estimates. In Table 2.3.1, formulas for $1-\alpha$ confidence limits are given for the population parameters μ, τ and p. These confidence intervals are only approximate since to be exact would require additional assumptions. However, for many situations, these confidence intervals are sufficiently valid to be useful. For further information on this subject, see Cochran Sec. 2.8, 2.15, and 3.6.

Table 2.3.1

Population Parameter	Lower Limit	Upper Limit
μ	$\hat{\mu} - t_{\alpha/2}\sqrt{\hat{V}(\hat{\mu})}$	$\hat{\mu} + t_{\alpha/2}\sqrt{\hat{V}(\hat{\mu})}$
τ	$\hat{\tau} - t_{\alpha/2}\sqrt{N^2\hat{V}(\hat{\mu})}$	$\hat{\tau} + t_{\alpha/2}\sqrt{N^2\hat{V}(\hat{\mu})}$
p	$\hat{p} - t_{\alpha/2}\sqrt{\hat{V}(\hat{p})}$	$\hat{p} + t_{\alpha/2}\sqrt{\hat{V}(\hat{p})}$

In Table 2.3.1, $t_{\alpha/2}$ is the upper $\alpha/2$ probability value of Student's t distribution with n-1 degrees of freedom.

2.3 Sample Size. When estimating a population parameter, one needs to know how large a sample to select so that his results will be useful. This has been discussed in some detail in Section 1.6. To determine the necessary sample size to give the desired results requires that the investigator specify a number d and a confidence coefficient $1-\alpha$ such that he wants the probability to be $1-\alpha$ that $\hat{\theta}$ does not deviate from θ by more than d units. In other words, for the population mean, total, and proportion, it is approximately correct that a $1-\alpha$ confidence interval on the parameter θ is $\hat{\theta} \pm t_{\alpha/2}\sqrt{\hat{V}(\hat{\theta})}$. Thus, if $t_{\alpha/2}\sqrt{\hat{V}(\hat{\theta})} \leq d$, then the requirements are met. If $\hat{V}(\hat{\theta})$ is replaced by $V(\hat{\theta})$, this can be set equal to d and the equation can be solved for n. The results for the population mean, total and proportion, are given below.

$$p: \quad n = \frac{t^2_{\alpha/2}\, pq}{d^2} \Bigg/ \left[1 + \frac{1}{N}\left(\frac{t^2_{\alpha/2}\, pq}{d^2} - 1\right)\right]$$

$$\mu: \quad n = \frac{t^2_{\alpha/2}\, S^2}{d^2} \Bigg/ \left[1 + \frac{t^2_{\alpha/2}\, S^2}{d^2 N}\right]$$

$$\tau: \quad n = \frac{t^2_{\alpha/2}\, N^2 S^2}{d^2 + t^2 N S^2}$$

To utilize these formulas, one substitutes values for the unknown parameters. These values may be obtained from a pilot study or from a previous study by some other method.

Chapter 3. SYSTEMATIC SAMPLING.

3.1 Introduction. Sometimes it is difficult to obtain a simple random sample. Or, an investigator may feel that estimates obtained from a simple random sample may not be adequate. For these reasons, other methods for selecting a sample from a population have been developed. One of these methods is systematic sampling. The idea is introduced with three examples.

Example 3.1.1. Suppose 2,000 cattle are to be trucked from summer range to a feedlot and the owner wants to examine the blood of the animals to determine if it contains a certain chemical associated with a pesticide that has been found in the water supply on the range. It is convenient to take the blood samples as the cattle are being loaded on trucks. To keep the operation running smoothly, it is desirable to get blood from every 40th animal just before it is loaded on the truck. It is clear that if blood is taken from every 40th animal that this is not a simple random sample. However, valid estimates of population parameters can be obtained and appropriate standard errors of the estimates can be computed if the first animal is selected at random and if certain assumptions can be made about the population. The method of selecting the sample (every 40th animal) is called a systematic (1-in-40) sampling method.

Example 3.1.2. Five years after a power line was constructed through 10 miles of an area, a study group must determine the amount of ground cover that has been restored on the right of way. A sample from 20 small areas along the line will be examined. A simple random sample is not desirable since that may result in observations that are not spread along the entire 10 miles. Since it is desirable to obtain a sample that is spread somewhat uniformly along the entire route, a systematic sample that allows observations to be taken at 1/2 mile intervals along the entire 10 miles would be very useful.

If the first sample is selected at random in the first 1/2 mile and if certain assumptions about the population are made, then valid estimates and standard errors can be obtained from the observations collected by this procedure.

Example 3.1.3. A forester plans to inventory a 40-acre stand of timber that is composed of a single species. He wants to estimate average merchantable volume per acre and total merchantable volume in the stand--both in board feet. The stand is located on a site that slopes uniformly downhill going from west to east. At all points on the site, there is a uniform increase in site quality as elevation decreases. Since merchantable volume increases as site quality increases, it is important that sample plots be located so that all site conditions are sampled.

The population described is an example of a population arranged in increasing order in that site quality (and hence volume) increases as elevation decreases. A systematic sample that would locate plots so that all site conditions were sampled may be superior to a simple random sample. One way of doing this would be to locate plots at fixed distances on lines running from east to west (or vice-versa). Several of these lines could be laid out in the stand.

From these examples one can observe that rather than select a simple random sample from a population, it may be more appropriate to select a sample that is more systematic. Two principal reasons for this are: 1) It may be convenient to select the sample in a systematic way as illustrated in Example 3.1.1. This convenience consideration may also simplify the procedure for selecting the sample and hence result in less observer errors. 2) It may be required that the observations be spread in a somewhat uniform manner over the entire population. In Example 3.1.2, a simple random sample could result

in all the sample observations being taken in the first five miles and the conclusions would not seem as convincing as if all ten miles were included.

From the examples, it is indicated that two things should be considered when selecting a systematic sample: 1) the total sample size denoted by n; 2) the sampling frequency (this is denoted by k). In Example 3.1.1, k=40 and hence every 40th animal is sampled; in Example 3.1.2, if the 10 miles is subdivided into 1/20 mile units, then the population size is 200. If a sample is selected every 1/2 mile, then k=10; i.e. every 10 units (1/20 mile long) equals 1/2 mile.

There is a restriction on the values of n and k for a given population size N. For given values of N and k a systematic sample of size n can be selected where n = N/k if N/k is an integer; if N/k is not an integer, then n is either the first integer less than N/k or the first integer greater than N/k.

To select a systematic sample, the population is organized in a certain order and then it is divided into groups of size k. This may be represented by:

$$\{Y_1, Y_2, \ldots, Y_k; Y_{k+1}, Y_{k+2}, Y_{2k}; \ldots;\}.$$

A number is selected at random from the k numbers 1, 2, ..., k (suppose it is the number t), then the sample consists of Y_t, Y_{t+k}, Y_{t+2k}, ...; i.e. the t-th number and every k-th unit thereafter.

Another way to view the sampling procedure is that the universe is divided into n groups of k items in each group (assume N=nk) and a sample of one item (the t-th) is taken from each group. It should be noted that only <u>one</u> item is selected at random from the universe and that is the "starting" item in the first group. After that item is selected (at random), then the remaining items from the universe that go into the sample (the t-th in each group)

are determined. The fact that only one item is chosen at random from the universe makes it impossible to obtain an unbiased estimate of the variance of estimates of population parameters unless additional assumptions are made about the arrangement of the population. This will be discussed in the next section. Below is the formal definition of a 1-in-k systematic sample.

Definition 3.1.1. 1-in-k Systematic Sample. The N items in a universe are numbered from 1 to N by some method and the first k items (k chosen by the investigator) are labeled group 1; the next k items are labeled group 2, etc. One number is selected at random from the set of numbers 1, 2, ..., k and represented by t. The Y_t and every k-th number thereafter in the population are the n sample numbers (the t-th number in each group). The sample obtained by this procedure is defined to be a 1-in-k systematic sample.

3.2 Point Estimates. The parameters of interest that are to be estimated are μ, τ and p. Unbiased estimates of these parameters exist using the observations from a systematic sample denoted by $y_1, y_2, \ldots, y_n$. The unbiased estimates are given in Table 3.2.1.

Table 3.2.1

Parameter	Unbiased Estimate
μ	$\hat{\mu} = \bar{y}$
τ	$\hat{\tau} = N\bar{y}$
p	$\hat{p} = \bar{y}$

To estimate p, the proportion of the universe that possesses a certain characteristic, the population numbers are either 0 or 1 (indicator numbers).

3.3 Confidence Interval Estimates. To compute (approximate) confidence intervals on the population parameters μ, τ, and p, unbiased estimators of the variances of the estimates in Table 3.2.1 are generally used. However, no unbiased estimates of the variances of the estimates in Table 3.2.1 are available unless it is known that the arrangement of the items of the universe satisfy certain conditions.

One such condition is that the arrangement of the population is random. In this case, estimates of $V(\hat{\mu})$, $V(\hat{\tau})$, and $V(\hat{p})$ that are unbiased when averaged over the appropriate population are available and are the same as for simple random samples. These are given in Table 3.3.1.

Table 3.3.1

Parameter	Unbiased Estimate of Variance of the Estimate
μ	$\hat{V}(\hat{\mu}) = (s^2/n)(1-f)$
τ	$\hat{V}(\hat{\tau}) = (N^2 s^2/n)(1-f)$
p	$\hat{V}(\hat{p}) = \frac{\hat{p}\hat{q}}{n-1}(1-f)$

If the items in the universe are arranged in random order, then for practical applications the systematic sample can be considered a simple random sample. The reason for using a systematic sample in this case is for convenience rather than for decreasing the variance of the estimates of population parameters.

To analyze the situation further, we consider three possible ways that the items in the universe might be arranged:

1) in random order,

2) in increasing (or decreasing) order,

3) in periodic order.

If the items are arranged as 1), then the estimates of the variances given in Table 3.3.1 are "unbiased." If the items are arranged as in 2), then the estimators given in Table 3.2.1 would have smaller variances under systematic sampling than simple random sampling. In addition, the estimators of variance given in Table 3.3.1 will overestimate the variance of systematic sampling estimators. If the items are arranged as in 3), then the variances of the estimates in Table 3.2.1 change depending on how the periodic nature of the population corresponds with the sampling interval. For further information on this subject, consult Chapter 8 of Cochran or Chapter 8 of Mendenhall.

Thus, we see that useful procedures do not exist for estimating the variances of point estimates of population parameters when using systematic sampling unless the universe satisfies 1). However, for situations 2) and 3), unbiased estimates of variances can be determined if repeated systematic sampling methods are used. (See Cochran, Chapter 8 or Mendenhall, Chapter 8).

Chapter 4. DATA FOR EXAMPLES

4.0 Introduction. The general purpose of this monograph is to provide comprehensive coverage of certain sets of sampling designs presently used or holding potential for use in natural resource inventories. These designs are multi-level in that they utilize various levels of sampling useful in estimation of population parameters. One through four levels are considered for multi-stage designs and multi-phase designs. This results in coverage of one- through four-stage and one- through four-phase designs.

The data termed "ground measurements" in this chapter are generally the only data used in the multi-stage designs. It is possible, however, to use other information when using ratio estimators in multi-stage sampling. Multi-stage sampling procedures -- including ratio estimators -- are presented in later chapters.

When using multi-phase designs, there is a certain "level" or set of information associated with each phase. Because we shall be describing procedures and presenting examples for one-phase through four-phase sampling designs, we shall be using four levels of information. These levels of information (data) are introduced in this chapter and are used for examples in the remainder of this monograph. The four levels used are Landsat imagery, high-altitude photography, low-altitude photography, and ground measurements.

4.1 Study Area. The area for which the data were developed essentially encompasses Grand County, Colorado. It is an area slightly over 1.1 million acres in size and is located west of the Continental Divide in north-central Colorado.

4.2 Landsat Imagery. Advances in remote sensing technology in the last several years have had important spin-offs in natural resource studies. Satellite

imagery, such as that provided by Landsat and used in this monograph, can provide a level of information on resource quality and quantities. Although this is considered crude in terms of what can be obtained from detailed measurements made on the ground, it still is useful, relatively inexpensive information. It holds promise of being extremely useful as auxiliary information in a sampling design framework.

The Landsat data used herein consist of 1,016,064 pixels (1.1 acre cells) arranged in a 1008 x 1008 grid. Each pixel is defined as belonging to one of seven area classes as determined by unsupervised classification of the entire grid. The population values consist of the ground measurements, or true values of each of the cells. The area classification of Landsat imagery is considered herein as auxiliary information which, together with sample values (and, at times, other auxiliary information), may be useful in the estimation of population parameters.

It is important to note here that it is not always efficient to use an entire population when simulating sampling. For example, if probability distributions or other criteria relating to population values are assumed, sampling simulation can often be conducted by generating only the sample values and disregarding those not included in the sample. This can often result in considerable savings in computer time and costs than when an entire population is placed in computer memory (or related storage devices) and subsequently sampled.

The entire universe of 1,016,064 pixels will be used for sampling examples in this monograph for two reasons. First, we hope that the population values and related auxiliary information will appear more realistic and comprehensible to the reader than if samples are generated by introduced criteria. Secondly,

although any comparisons of examples for different designs lead to conclusions related only to the population used in the examples, selection criteria under systematic sampling almost necessitate knowledge of the entire population.

The seven area classes used for the satellite imagery are:

1. Conifers - areas at least 25% stocked by trees with at least 50% of the tree crown closure in conifers (5-acre minimum).
2. Sage - nonforest land with the predominance of the vegetative cover in sagebrush.
3. Hardwoods - areas at least 25% stocked by trees with at least 51% of the tree crown closure in hardwoods (5-acre minimum).
4. Meadow - nonforest land with the predominance of the vegetative cover in cultivated grass meadow or pasture. Also includes cropland.
5. Shrub - nonforest land with the predominance of the vegetative cover in brush other than sagebrush.
6. Grass - nonforest land with the predominance of the vegetative cover in native uncultivated grass or in barren ground.
7. Water - areas over 5 acres in size and over 200 feet wide that appear to be permanent standing year-round water.

4.3 High-Altitude Photography. High-altitude aerial photography, just as has been the case with satellite imagery, has enlarged the possible approaches to natural resources sampling. Photos with scales of 1:60,000 or smaller have proven useful in delineation -- at least approximate delineation -- of vegetation types and other broad classes. As with satellite imagery, high-altitude photography is usually generated for purposes other than -- or in addition to -- natural resources studies. Thus, it is often relatively inexpensive to use if

it can be incorporated into a sampling design using various levels of information. The photography used for examples in this monograph is 1:130,000 color infrared.

It would be costly to classify all one million cells into the seven classes used for the satellite imagery. Because the intent of using information at four levels is for illustrative examples, it was not considered feasible to classify all cells. Instead, a "confusion" matrix was developed from a sample of cells. In Table 4.3.1, each column represents an area classification from the satellite imagery; each row represents a classification from the high-altitute photography. Each entry in the table is a probability of classification. Thus, for example, if a cell was classified as area class 2 on the satellite imagery, the probability is 0.7329 that the cell would be called class 2 by photo interpretation of the high-altitude photography.

Table 4.3.1

Landsat/High-Altitude Confusion Matrix

High-Altitude Photography Area Class	Landsat Area Class						
	1	2	3	4	5	6	7
1	0.8077	0.0	0.0233	0.0	0.0	0.0160	0.0440
2	.0656	.7329	.0651	.0167	.2949	.0856	.0
3	.0746	.0685	.6791	.0500	.1026	.0107	.0
4	.0	.0616	.1442	.9208	.0	.1230	.0
5	.0317	.0754	.0651	.0	.5513	.0321	.0
6	.0	.0616	.0232	.0125	.0512	.7326	.0
7	.0204	.0	.0	.0	.0	.0	.9560

Using Table 4.3.1, which was developed from 1400 selected cells, each cell was assigned an area class for the level of information representing high-altitude photography.

At this point, the reader may question why such a large population is being used. Wouldn't it be simpler to use a very small population with complete interpretation?

It would certainly be simpler, but we did not consider it to be a practical approach if we were to present the examples as being realistic counterparts of those encountered in actual sampling studies. Consider a case where we want to demonstrate four-stage sampling. At the fourth stage of sampling we have the population values; in this case, a population value is a measurement made on a 1.1 acre cell. The third level of information represents a grouping of cells. Here it might be realistic to consider approximately 10 acres (acutally 3 x 3 = 9 cells or 9.9 acres) as a third-level unit. Moving upwards to the second level, we could work with approximately a section of land (actually 8 x 8 = 64 9.9 acre blocks = 633.6 acres). At the first level, we could then be working with townships (actually 6 x 6 = 36 633.6 acre blocks = 22,809.6 acres). In fact, with 1,016,064 pixels, we have only 49 first-level or primary units in the population. We, therefore, consider the population used in these examples as a relatively small population; anything smaller would not be realistic.

4.4 Low-Altitude Photography. Aerial photographs with scales of 1:25,000 and larger have been used for several decades in natural resources sampling. Here, as with the high-altitude photography, a probability matrix (Table 4.4.1) was used to generate values for the cells. The photography used to develop the table entries is 1:25,000 color infrared. A total of 224 selected cells was used to develop the table entries.

Table 4.4.1

High-Altitude Photography/Low-Altitude Photography

Confusion Matrix

Low-Altitude Photography Area Class	High-Altitude Photography Area Class						
	1	2	3	4	5	6	7
1	0.7174	0.0	0.0588	0.0	0.0	0.0	0.0
2	0.0	0.5435	0.0588	0.0	0.1915	0.1667	0.0
3	0.2826	0.0	0.8824	0.0455	0.0	0.0	0.0
4	0.0	0.0	0.0	0.8636	0.2553	0.1667	0.0
5	0.0	0.2174	0.0	0.0455	0.3830	0.1111	0.0
6	0.0	0.2391	0.0	0.0454	0.1702	0.5000	0.0
7	0.0	0.0	0.0	0.0	0.0	0.0555	1.000

4.5 Ground Measurements. The ground measurements are the population values. We shall be dealing with three items of interest in our examples. The first is land area by classes. We shall use the same classes used at the other three information levels. Secondly, we shall be estimating merchantable timber volume per pixel in cubic feet; thus we need a measure of volume per acre for each of the cells. Lastly, we shall estimate herbage in pounds per pixel.

As with the two levels of photography, it would be ideal to have known values for each cell. However, practicality dictated that these data be generated with Monte Carlo techniques also. Information derived from selected cells within the study area is presented in Table 4.5.1 which was subsequently used to assign values for area classification to each of the 1.1 million cells.

Table 4.5.1

Low Altitude Photography/Ground Measurement
Confusion Matrix

Ground Measurement Area Class	Low-Altitude Photography Area Class						
	1	2	3	4	5	6	7
1	0.9000	0.0	0.2500	0.0	0.0	0.0	0.0
2	0.0	0.8947	0.0	0.0857	0.4705	0.3636	0.0
3	0.0500	0.0	0.7000	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.8286	0.0882	0.0606	0.0
5	0.0300	0.0264	0.0500	0.0286	0.1176	0.0303	0.0
6	0.0200	0.0789	0.0	0.0571	0.3236	0.5455	0.0333
7	0.0	0.0	0.0	0.0	0.0	0.0	0.9667

The timber volume and herbage values were generated by a Monte Carlo technique using means and standard deviations supplied by the U. S. Forest Service. Assignment of values was based on the area classification (ground measurement) derived previously. Distributions were assumed to be truncated

normal (truncated at zero) with mean μ and variance σ^2.

Table 4.5.2
Timber Volume and Herbage Values by Area Class

Area Class	μ [1]	σ [1]
1	2453	1847.4
2	3300	1320.0
3	1073	788.0
4	2200	660.0
5	3850	2115.8
6	825	330 .0
7	0	0.0

The final result is a complete tabulation of population values for three variables (actually we then have three populations)as well as auxiliary information at three other levels. The composite information, although not the exact information we would have if we could completely enumerate the study area, represents something very close to a real situation. And we believe the examples given throughout this manual are typical of what would result from actual sampling studies conducted in an area similar to the study area.

The population values which were derived from the procedures described in these sections are summarized below in terms of parameters which will be estimated in the examples that follow in the remainder of this monograph:

μ(timber volume) = 1,043.7 cubic fee per pixel

σ(timber volume) = 1,616.6 cubic feet

μ(herbage) = 1,225.1 pounds per pixel

σ(herbage) = 1,901.8 pounds

p(forest) = proportion of forest land = 0.4888

1/ For classes 1 and 3 μ and σ are in cubic feet of timber volume per pixel. For other classes, μ and σ are in pounds of herbage per pixel. If timber volume was present, μ and σ for herbage are zero; if herbage was present, μ and σ for timber volume are zero.

4.6 Examples. Simple random sampling and systematic sampling will be used here with the population just described. In these examples and those in subsequent chapters, the reader may wish to compare the results with the population values given previously.

Example 4.6.1 Given a universe of N = 1,016,064 1.1 - acre pixels, a simple random sample (without replacement) of n = 1008 pixels was selected to estimate average timber volume per pixel ($\hat{\mu}$ timber volume), average herbage per pixel ($\hat{\mu}$ herbage), and the proportion of forest area ($\hat{p}$ forest). The results were:

	$\hat{\theta}$	$\sqrt{\hat{V}(\hat{\theta})}$	$100\sqrt{\hat{V}(\hat{\theta})}/\hat{\theta}$
timber volume, $\hat{\theta} = \hat{\mu}$ timber volume	999.1	44.7	4.5%
herbage, $\hat{\theta} = \hat{\mu}$ herbage	1243.1	51.0	4.1%
forest area, $\hat{\theta} = \hat{p}$ forest	0.4752	0.0157	3.3%

Example 4.6.2 Given a universe of N = 1,016,064 1.1 - acre pixels, a 1 in 1200 systematic sample of n = 846 pixels was selected to estimate average timber volume per pixel ($\hat{\mu}$ timber volume), average herbage per pixel ($\hat{\mu}$ herbage), and the proportion of forest area ($\hat{p}$ forest). The results were:

	$\hat{\theta}$	$\sqrt{\hat{V}(\hat{\theta})}$ [1/]	$100\sqrt{\hat{V}(\hat{\theta})}/\hat{\theta}$
timber volume, $\hat{\theta} = \hat{\mu}$ timber volume	1067.6	50.9	4.8%
herbage, $\hat{\theta} = \hat{\mu}$ herbage	1229.5	57.3	4.7%
forest area, $\hat{\theta} = \hat{p}$ forest	0.4917	0.0172	3.5%

Example 4.6.3 Given a universe of N = 1,016,064 1.1 - acre pixels, a 1 in 800 systematic sample of n=1270 pixels was selected to estimate average timber volume per pixel ($\hat{\mu}$ timber volume), average herbage per pixel ($\hat{\mu}$ herbage), and the proportion of forest area ($\hat{p}$ forest). The results were:

	$\hat{\theta}$	$\sqrt{\hat{V}(\hat{\theta})}$ [1/]	$100\sqrt{\hat{V}(\hat{\theta})}/\hat{\theta}$
timber volume, $\hat{\theta} = \hat{\mu}$ timber volume	1074.5	41.7	3.9%
herbage, $\hat{\theta} = \hat{\mu}$ herbage	1199.0	45.2	3.8%
forest area, $\hat{\theta} = \hat{p}$ forest	0.5031	0.0140	2.8%

[1/] Using simple random sampling formulae.

PART II

MULTISTAGE DESIGNS

Chapter 5. ONE- AND TWO-STAGE SAMPLING.

5.1 Introduction. In many applications in natural-resource studies, it is quite costly or impractical to select a sample of items by using simple random sampling or systematic sampling. Two common situations when these two sampling designs are not attractive are 1) When the distance between sampled items is so vast that the time and cost of covering the entire area is prohibitive, or 2) When it is virtually impossible to identify each item in the universe. In these situations, it is often useful to divide the items into N nonoverlapping groups and select a simple random sample of n ($n \geq 2$) of the N groups (often called clusters) and measure some or all of the items in the clusters selected. If all of the items in each selected cluster are measured the design is called a single-stage (cluster) sampling design; if all items in each selected cluster are not measured (but rather a sample from each cluster is measured) the design is called a two-stage sampling design.

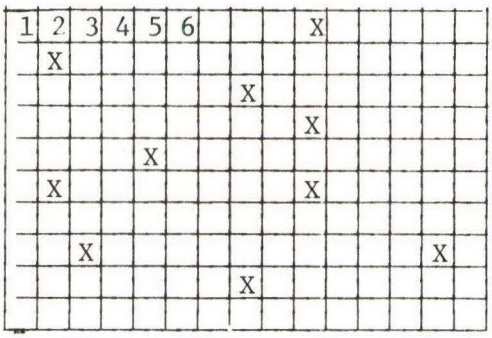
In Section 5.2, a single-stage design is defined and estimation procedures are discussed. Two-stage sampling designs are discussed in Section 5.3. In Chapters 6 and 7, respectively, three-stage and four-stage designs are discussed.

5.2 One-Stage Sampling Design. This design is often referred to as 1) a cluster sampling design, 2) a single-stage sampling design or, 3) a one-factor nested design.

The topic to be discussed in this section will be introduced with an example.

Example 5.2.1 A study group wants to determine the average age of mature trees in a 3,000 acre tract. It is decided to divide the tract into 150 acres of 20

acres each and use cluster sampling. The area is illustrated below.



The 150 areas are numbered from 1 to 150 by some method of numbering and a simple random sample of 10 numbers is chosen. The numbers selected are the ten tracts that will be thoroughly measured; i.e. the age of each mature tree in these ten tracts will be determined.

The definition of a single-stage sampling design is given below.

Definition 5.2.1 Single-Stage Sampling Design.

If the items in a universe are divided into N groups, often called clusters, such that each item is contained in exactly one cluster, and if all items in each sampled cluster are observed, the sampling design is defined to be a single-stage sampling design.

Note The definition does not state how the sample clusters are selected. We assume here that the clusters are obtained by simple random sampling.

The following notation will be used throughout this section.

N = number of clusters in the universe

M_i = number of items in the i^{th} cluster for $i = 1, \ldots, N$.

$M_0 = \sum_{i=1}^{N} M_i$ = total number of items in the universe

$\bar{M} = M_0/N$ = average number of items per cluster

$Y_{i.}$ = total of measurements of all items in the ith cluster in the universe; i = 1, 2, ..., N

n = number of clusters in the simple random sample of clusters

$y_{i.}$ = total of measurements of all items in the ith cluster in the sample; i = 1, 2, ..., n.

$\bar{m} = \frac{1}{n} \sum_{i=1}^{n} m_i$ = average cluster size for those clusters in the sample.

The quantities above that are known are N, m_i (for the clusters in the sample), n, $\bar{m}$, $y_{i.}$; i = 1, 2, ..., n.

Estimators and the estimated variances of the estimators of μ, τ, and p are given in Table 5.2.1.

Table 5.2.1

Population Parameter	Point Estimate	Variance Estimate
μ	$\hat{\mu} = \sum_{i=1}^{n} y_{i.} / \sum_{i=1}^{n} m_i$	$\hat{V}[\hat{\mu}] = \frac{N-n}{N} \sum_{i=1}^{n} \frac{(y_{i.} - \hat{\mu} m_i)^2}{\bar{M}^2 \, n(n-1)}$
τ	$\hat{\tau} = \frac{N}{n} \sum_{i=1}^{n} y_{i.} = N\bar{y}$	$\hat{V}[\hat{\tau}] = \frac{N-n}{N} N^2 \sum_{i=1}^{n} \frac{(y_{i.} - \bar{y})^2}{n(n-1)}$
p	$\hat{p} = \sum_{i=1}^{n} m_i p_i / \Sigma m_i$	$\hat{V}[\hat{p}] = \frac{N-n}{N} \sum_{i=1}^{n} \frac{m_i^2 (p_i - \hat{p})^2}{\bar{M}^2 \, n(n-1)}$

In Table 5.2.1, $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$; p_i is the proportion of items in the ith sampled cluster that has the characteristic under study.

Example 5.2.1 Given a universe of N = 112,896 9.9-acre blocks, 336 blocks on a side, each block containing nine 1.1-acre pixels, three pixels on a side, a simple random sample of n = 112 blocks was selected. Within each selected block, all nine pixels were measured, resulting in a single-stage sample of 1008 pixels to estimate average timber volume per pixel ($\hat{\mu}$ timber volume),

average herbage per pixel ($\hat{\mu}$ herbage), and the proportion of forest area ($\hat{p}$ forest). The results were:

	$\hat{\theta}$	$\sqrt{\hat{V}(\hat{\theta})}$	$100\sqrt{\hat{V}(\hat{\theta})}/\hat{\theta}$
timber volume, $\hat{\theta} = \hat{\mu}$ timber volume	1016.7	81.4	8.0%
herbage, $\hat{\theta} = \hat{\mu}$ herbage	1301.2	87.7	6.7%
forest area, $\hat{\theta} = \hat{p}$ forest	0.4762	0.0331	7.0%

5.3 Two-Stage Sampling Designs. One extension of single-stage sampling is the two-stage sample design. It is useful when it is costly, impractical, or impossible to measure every item in the sampled clusters selected in single-stage designs. For instance, in Example 5.2.1, it may be very costly to measure the age of each mature tree in each sampled cluster, so a sample of the mature trees within each cluster is selected and their ages measured. The clusters are called the primary sampling units (or first stage units), and the items within the clusters are called the secondary sampling units (or the subsample units). Since the sample is selected in two stages, the design is appropriately termed a two-stage sampling design. This topic will be introduced with examples. Results of the examples given are based on the unbiased estimators given later in this chapter.

Example 5.3.1 Given a universe of N = 49 townships, seven townships on a side, each township containing M = 20,736 1.1-acre pixels, 144 pixels on a side, a simple random sample of n = 9 townships was selected. Within each selected township, a simple random sample of m = 112 pixels was selected, resulting in a two-stage sample of 1008 pixels to estimate average timber volume per pixel ($\hat{\mu}$ timber volume), average herbage per pixel ($\hat{\mu}$ herbage), and the proportion of forest area ($\hat{p}$ forest). The results were:

	$\hat{\theta}$	$\sqrt{\hat{V}(\hat{\theta})}$	$100\sqrt{\hat{V}(\hat{\theta})}/\hat{\theta}$
timber volume, $\hat{\theta} = \hat{\mu}$ timber volume	883.6	132.1	15.0%
herbage, $\hat{\theta} = \hat{\mu}$ herbage	1532.8	141.8	9.3%
forest area, $\hat{\theta} = \hat{p}$ forest	0.3790	0.0549	14.5%

Example 5.3.2 Given a universe of N = 1764 sections, 42 sections on a side, each section containing M = 576 1.1-acre pixels, 24 pixels on a side, a simple random sample of n = 42 sections was selected. Within each selected section, a simple random sample of m = 24 pixels was selected, resulting in a two-stage sample of 1008 pixels to estimate average timber volume per pixel ($\hat{\mu}$ timber volume), average herbage per pixel ($\hat{\mu}$ herbage), and the proportion of forest area. The results were:

	$\hat{\theta}$	$\sqrt{\hat{V}(\hat{\theta})}$	$100\sqrt{\hat{V}(\hat{\theta})}/\hat{\theta}$
timber volume, $\hat{\theta} = \hat{\mu}$ timber volume	884.0	103.7	11.7%
herbage, $\hat{\theta} = \hat{\mu}$ herbage	1364.0	113.5	8.3%
forest area, $\hat{\theta} = \hat{p}$ forest	0.4246	0.0451	10.6%

Example 5.3.3 Given a universe of N = 112,896 9.9-acre blocks, 336 blocks on a side, each block containing M = 9 1.1-acre pixels, three pixels on a side, a simple random sample of n = 504 blocks was selected. Within each selected block, a simple random sample of m = 2 pixels was selected, resulting in a two-stage sample of 1008 pixels to estimate average timber volume ($\hat{\mu}$ timber volume), average herbage per pixel ($\hat{\mu}$ herbage), and the proportion of forest area. The results were:

	$\hat{\theta}$	$\sqrt{\hat{V}(\hat{\theta})}$	$100\sqrt{\hat{V}(\hat{\theta})}/\hat{\theta}$
timber volume, $\hat{\theta} = \hat{\mu}$ timber volume	994.5	52.1	5.2%
herbage, $\hat{\theta} = \hat{\mu}$ herbage	1221.7	54.1	4.4%
forest area, $\hat{\theta} = \hat{p}$ forest	0.4500	0.0185	4.1%

Two-stage sampling is commonly used in natural resource inventories as an alternative to one-stage sampling if the items in the sampled clusters are widely scattered such that travel from one item to another involves significant cost. The efficiency of two-stage sampling relative to one-stage sampling or simple random sampling depends on the characteristics of the specific population of interest. The distance between items (and thus the total travel cost), the number of items in the sampled clusters (and thus the cost of measuring the entire cluster), and the variation among and within clusters are all contributing factors to a comparison of the approaches.

It is not surprising to note that two-stage sampling has proved efficient in timber inventories of jungle areas of developing countries. Remote locations and difficult access in such areas cause extremely high travel costs. Thus, for a fixed allowable error, it is usually less costly to select several clusters and measure some items within each than it would be to measure the number of items under simple random or one-stage sampling required for the same allowable error.

While the data used for examples in this monograph do not necessarily exhibit the characteristics which suggest efficiency under two-stage sampling, they are at least useful in demonstrating the procedures involved. The formal definition of a two-stage sampling design is given below.

Definition 5.3.1. Two-Stage Sampling Design. Consider a universe that is divided into N groups so that each item in the universe is in exactly one group. The N groups are called first-stage (primary) sampling units (clusters). If each first-stage unit is divided into second-stage (secondary) units so that

the ith first-stage unit contains M_i sub-units (items), and so that the jth sub-unit (for j = 1, ..., M_i) in the ith primary unit (for i = 1, ..., N) is contained in exactly one primary unit, the resulting design is called a two-stage sampling design.

Note. The definition does not state how the primary units and the secondary units (sub-units) are selected for observation. For example, n primary units may be selected from the N universe units by simple random sampling. Then from the ith selected primary unit, a simple random sample of m_i secondary units may be selected for i = 1, ..., n; or a systematic sample of m_i secondary units may be selected from the ith sample primary unit. Other methods of selecting the primary sample and the subsample could be used. Some of these procedures and the resulting estimates of appropriate population parameters will be discussed later.

The following notation will be used in this section.

N = number of primary units in the universe.

M_i = number of secondary units in the ith primary unit for i = 1, ..., N.

$M_0 = \sum_{i=1}^{N} M_i$ = total number of secondary units in the universe

$\bar{M} = M_0/N$ = average number of secondary units per primary unit in the universe.

Y_{ij} = population value of character under study in the jth secondary unit of the ith primary unit; j = 1, ..., M_i; i = 1, ..., N.

$\mu_i = \frac{1}{M_i} \sum_{j=1}^{M_i} Y_{ij}$ = average of population values of ith primary unit for

i = 1, ..., N.

$\mu = \sum_{i=1}^{N} M_i \mu_i / \sum_{k=1}^{N} M_k = \sum_{i=1}^{N} \sum_{j=1}^{M_i} Y_{ij} / \sum_{k=1}^{N} M_k = Y../M_0$ = mean of all population values

$$Y.. = \sum_{i=1}^{N} \sum_{j=1}^{M_i} Y_{ij} = \sum_{i=1}^{N} Y_{i.} = \text{total of all population values}$$

$$S_{2i}^2 = \frac{1}{M_i - 1} \sum_{j=1}^{M_i} (Y_{ij} - \mu_i)^2$$

n = number of primary units selected in the sample

m_i = number of secondary units selected in the sample from the ith sample primary unit.

y_{ij} = value (observed measurement) of the jth sampled secondary unit in the ith sampled primary unit; $j = 1, \ldots, m_i$; $i = 1, \ldots, n$.

$$\bar{y}_{i.} = \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij} = \text{sample mean of ith sample primary unit for } i = 1, \ldots, n$$

$$\bar{y}.. = \sum_{i=1}^{n} m_i \bar{y}_{i.} \Big/ \sum_{j=1}^{n} m_j = \text{average of all sample values}$$

$$s_1^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\bar{y}_{i.} - \bar{y}..)^2 = s^2(\bar{y}_{i.} \mid i,n)$$

$$s_{2i}^2 = \frac{1}{m_i - 1} \sum_{j=1}^{m_i} (y_{ij} - \bar{y}_{i.})^2 = s^2(\bar{y}_{i.} \mid j, m_i) = \text{sample variance of the ith sample primary unit for } i = 1, \ldots n.$$

$$y.. = \sum_{i=1}^{n} y_{i.} = \sum_{i=1}^{n} \sum_{j=1}^{m_i} y_{ij} = \text{total of all sample values}$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_{i.} = \text{average per primary unit}$$

$$m_0 = \sum_{i=1}^{n} m_i = \text{total number of subunits sampled}$$

$$f_1 = \frac{n}{N};\ f_2 = \frac{m}{M};\ f_{2i} = \frac{m_i}{M_i}$$

If each primary unit has the same number M (known) of secondary units, and if an equal number m of secondary units is sampled in each sampled primary unit, the formulas for estimators of μ, τ, and p and for the estimators of the variances are given in Table 5.3.1.

Table 5.3.1

Population Parameter	Point Estimate	Variance Estimate
μ	$\hat{\mu} = \bar{y}..$	$\hat{V}(\hat{\mu}) = \frac{1 - f_1}{n} s_1^2 + \frac{f_1(1 - f_2)}{mn} s_2^2$ where $s_1^2 = \sum(\bar{y}_{i.} - \hat{\mu})^2/(n-1)$ and $s_2^2 = \sum_i^n \sum_j^m \frac{(y_{ij} - \bar{y}_{i.})^2}{n(m-1)}$
τ	$\hat{\tau} = M_0\bar{y}..$	$\hat{V}(\hat{\tau}) = (NM)^2\hat{V}(\hat{\mu})$
p[1]	$\hat{p} = \bar{y}..$	$\hat{V}(\hat{p}) = \frac{1 - f_1}{n} s_2^2 + \frac{f_1(1-f_2)}{mn} s_2^2$ where $s_1^2 = \sum(\bar{y}_{i.} - \hat{\mu})^2/(n-1)$ and $s_2^2 = \sum_i^n \sum_j^m \frac{(y_{ij} - \bar{y}_{i.})^2}{n(m-1)}$

Note. The estimates of μ, τ, and p are unbiased estimates. Also, the variance estimates are unbiased estimates of the respective population variances.

In some situations, an auxiliary variable which we denote by X_{ij} is available on each sampled item and often this variable can be useful along with the principal variable Y_{ij} in estimating the mean μ or total τ of a population. The resulting estimate is called a "ratio" estimate and the appropriate formulas are given in Table 5.3.2.

[1]For the proportion $y_{ij} = 1$ if the secondary unit has the characteristic under study and $y_{ij} = 0$ if it does not.

Table 5.3.2

Population Parameter	Point Estimate	Variance Estimate
μ	$\hat{\mu} = (\bar{y}../\bar{x}..)\mu_x$	$\hat{V}(\hat{\mu}) = \frac{(1 - f_1)}{n} \sum_{i=1}^{n} \frac{(\bar{y}_{i.} - \hat{R}\bar{x}_{i.})^2}{n - 1} + \frac{f_1}{n^2} \sum_{i=1}^{n} \frac{(1 - f_{2i})}{m} s^2_{d2i}$ where $\hat{R} = \Sigma\bar{y}_{i.}/\Sigma\bar{x}_{i.}$ and $s^2_{d2i} = \frac{\sum_{j=1}^{m} (dij - \bar{d}i.)}{m - 1}$ where $dij = y_{ij} - \hat{R}x_{ij}$
τ	$\hat{\tau} = (\bar{y}../\bar{x}..)\tau_x$	$\hat{V}(\hat{\tau}) = N^2M^2 \hat{V}(\hat{\mu})$, where $\hat{V}(\hat{\mu})$ is given by Equation (5.3.18)

Note. The point estimates of μ and τ in Table 5.3.2 are not unbiased estimates; the variance estimates are not unbiased estimates.

In the above estimates, the reader will notice that in various formulas it is required to know certain population values such as μ_x, τ_x, M_0. Also, the formulas in Tables 5.3.1 and 5.3.2 are valid only for the case of equal number of secondary units in each primary unit (i.e. $M_1 = M_2 = \ldots = M_N = M$) and when an equal number of secondary units ($m_1 = m_2 = \ldots = m_n = m$) have been selected from each sampled primary unit. Also, the primary units and secondary units have each been selected by simple random sampling.

Often it is not possible to partition the universe into primary units which each contain the same number of secondary units (i.e. $M_i \neq M$). Also, it is not always possible or desirable to sample the same number of secondary units from each primary unit (i.e. $m_i \neq m$). It is not always desirable to select the primary and secondary units by simple random sampling. These various situations are outlined below and the appropriate point estimates of μ, τ, and p, and the estimated variances are given in Table 5.3.3 that follows. Symbols given under equation numbers refer to Lemmas (L) used in development of the equations or chapter numbers in Cochran (C) or Sukhatme (S) where derivations are given. The lemmas appear in the appendix.

Formulas for Two-Stage Design.

(5.3.1) $$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \bar{y}_{i.}, \text{ where } \bar{y}_{i.} = \frac{1}{m} \sum_{j=1}^{m} y_{ij}.$$

(5.3.2) $$\hat{V}(\hat{\mu}) = \frac{1 - f_1}{n} s_1^2 + \frac{f_1(1 - f_2)}{mn} s_2^2$$

where $s_1^2 = \sum_{i=1}^{n} (\bar{y}_{i.} - \hat{\mu})^2/(n-1)$

and $s_2^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(y_{ij} - \bar{y}_{i.})^2}{n(m-1)}$

Table 5.3.3

Primary Unit Size / Sample Size	Method of Selection		Type of Estimator	Equation #			
				MEAN & PROPORTION		TOTAL	
	Primary Units	Secondary Units		Point Estimate	Variance Estimate	Point Estimate	Variance Estimate
M_i/m_i	SRS	SRS	U	5.3.1 5.3.5 5.3.9 5.3.13	5.3.2 5.3.6 5.3.10 5.3.14	5.3.3 5.3.7 5.3.11 5.3.15	5.3.4 5.3.8 5.3.12 5.3.16
M_i/m_i	SRS	SRS	R	5.3.17 5.3.21 5.3.25 5.3.29	5.3.18 5.3.22 5.3.26 5.3.30	5.3.19 5.3.23 5.3.27 5.3.31	5.3.20 5.3.24 5.3.28 5.3.32
M_i/m_i	SRS	SYST	U	5.3.1 5.3.5 5.3.9 5.3.13	5.3.33 & 5.3.34 5.3.33 & 5.3.34 5.3.33 & 5.3.34 5.3.33 & 5.3.34	5.3.3 5.3.7 5.3.11 5.3.15	5.3.35 & 5.3.36 5.3.35 & 5.3.36 5.3.35 & 5.3.36 5.3.35 & 5.3.36
M_i/m_i	SRS	SYST	R	5.3.17 5.3.21 5.3.25 5.3.29	1/	5.3.19 5.3.23 5.3.27 5.3.31	1/ 1/
M_i/m_i	SYST	SRS	U	5.3.1 5.3.5 5.3.9 5.3.13	1/	5.3.5 5.3.7 5.3.11 5.3.15	1/
M_i/m_i	SYST	SRS	R	5.3.17 5.3.21 5.3.25 5.3.29	1/	5.3.19 5.3.23 5.3.27 5.3.31	1/
M_i/m_i	SYST	SYST	U	5.3.1 5.3.5 5.3.9 5.3.13	1/	5.3.5 5.3.7 5.3.11 5.3.15	1/
M_i/m_i	SYST	SYST	R	5.3.17 5.3.21 5.3.25 5.3.29	1/	5.3.19 5.3.23 5.3.27 5.3.31	1/
M_i/m_i	Probability Proportional to P_i: with replacement	SYST	U	5.3.37 5.3.37 5.3.37 5.3.37	5.3.38 5.3.38 5.3.38 5.3.38	5.3.39 5.3.39 5.3.39 5.3.39	5.3.40 5.3.40 5.3.40 5.3.40

1/ Generally accepted methods not available without making further assumptions.

2/ The symbols U and R denote unbiased and ratio estimates respectively.

(5.3.3) $$\hat{\tau} = \frac{NM}{n} \sum_{i=1}^{n} \bar{y}_{i.} = NM\hat{\mu}, \text{ where } \bar{y}_{i.} = \frac{1}{m} \sum_{j=1}^{m} y_{ij}$$

(5.3.4) $$\hat{V}(\hat{\tau}) = (NM)^2 \hat{V}(\hat{\mu})$$

where $\hat{V}(\hat{\mu})$ is given by Equation (5.3.2).

(5.3.5) C 10 $$\hat{\mu} = \frac{N}{nM_0} \sum_{i=1}^{n} M_i \bar{y}_{i.} \text{ where } \bar{y}_{i.} = \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij}$$

(5.3.6) C10 $$\hat{V}(\hat{\mu}) = \frac{1 - f_1}{n(\bar{M}.)^2} \sum_{i=1}^{n} \frac{(M_i \bar{y}_{i.} - \hat{\bar{Y}})^2}{n - 1}$$

$$+ \frac{f_1}{n^2(\bar{M}.)^2} \sum_{i=1}^{n} \frac{M_i^2(1 - f_{2i})s_{2i}^2}{m_i}$$

where $\hat{\bar{Y}} = \sum_{i=1}^{n} M_i \bar{y}_{i.}/n$ and $s_{2i}^2 = \frac{\sum_{j=1}^{m_i} (y_{ij} - \bar{y}_{i.})^2}{m_i - 1}$

(5.3.7) C10 $$\hat{\tau} = \frac{N}{n} \sum_{i=1}^{n} M_i \bar{y}_{i.}$$

(5.3.8) C10 $\hat{V}(\hat{\tau}) = M_0^2 \hat{V}(\hat{\mu})$, where $\hat{V}(\hat{\mu})$ is given by equation (5.3.6).

(5.3.9) $$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \bar{y}_{i.}$$

(5.3.10) $$\hat{V}(\hat{\mu}) = \frac{1 - f_1}{n} \sum_{i=1}^{n} (\bar{y}_{i.} - \hat{\mu})^2/(n - 1)$$

$$+ \frac{f_1}{n^2} \sum_{i=1}^{n} \frac{(1 - f_{2i})}{m_i} s_{2i}^2$$

(5.3.11) $$\hat{\tau} = \frac{NM}{n} \sum_{i=1}^{n} \bar{y}_{i.}$$

(5.3.12) $\hat{V}(\hat{\tau}) = N^2M^2 \hat{V}(\hat{\mu})$, where $\hat{V}(\hat{\mu})$ is given by Equation (5.3.10)

(5.3.13) $$\hat{\mu} = \frac{N}{nM_0} \sum_{i=1}^{n} M_i \bar{y}_{i.} \text{ where } \bar{y}_{i.} = \frac{1}{m} \sum_{j=1}^{m} y_{ij}$$

(5.3.14) $$\hat{V}(\hat{\mu}) = \frac{1 - f_1}{n(\bar{M}.)^2} \sum_{i=1}^{n} (M_i \bar{y}_{i.} - \hat{\bar{Y}})^2/n-1$$

$$+ \frac{f_1}{n^2(\bar{M}.)^2} \sum_{i=1}^{n} \frac{M_i^2(1 - f_{2i})}{m} s_{2i}^2$$

where $\hat{\bar{Y}} = \sum_{i=1}^{n} M_i \bar{y}_{i.}/n$ and $s_{2i}^2 = \frac{\sum_{j=1}^{m} (y_{ij} - \bar{y}_{i.})^2}{m - 1}$

(5.3.15) $$\hat{\tau} = \frac{N}{n} \sum_{i=1}^{n} M_i \bar{y}_{i.}$$

(5.3.16) $\hat{V}(\hat{\tau}) = M_0^2 \hat{V}(\hat{\mu})$, where $\hat{V}(\hat{\mu})$ is given by Equation (5.3.14).

(5.3.17) $$\hat{\mu} = \left(\sum_{i=1}^{n} y_{i.} / \sum_{i=1}^{n} x_{i.} \right) \mu_x$$

(5.3.18) $$\hat{V}(\hat{\mu}) = \frac{(1 - f_1)}{n} \sum_{i=1}^{n} \frac{(\bar{y}_{i.} - \hat{R}\bar{x}_{i.})^2}{n - 1} + \frac{f_1}{n^2} \sum_{i=1}^{n} \frac{(1 - f_{2i})}{m} s_{d2i}^2$$

where $\hat{R} = \sum_{i=1}^{n} \bar{y}_{i.} / \sum_{i=1}^{n} \bar{x}_{i.}$

$$s_{d2i} = \frac{\sum_{j=1}^{m} (dij - \bar{d}i.)}{m - 1} \quad \text{where } dij = y_{ij} - \hat{R}x_{ij}$$

(5.3.19) $\hat{\tau} = NM(\hat{\mu})$, where $\hat{\mu}$ is given by Equation (5.3.17).

(5.3.20) $\hat{V}(\hat{\tau}) = N^2M^2 \hat{V}(\hat{\mu})$, where $\hat{V}(\hat{\mu})$ is given by Equation (5.3.18).

(5.3.21) C10 $$\hat{\mu} = \mu_x \left(\sum_{i=1}^{n} M_i \bar{y}_i / \sum_{i=1}^{n} M_i \bar{x}_i \right)$$

(5.3.22) C10 $$\hat{V}(\hat{\mu}) = \frac{1 - f_1}{n} \sum_{i=1}^{n} \frac{M_i^2(\bar{y}_{i.} - \hat{R}\bar{x}_{i.})^2}{n - 1} + \frac{f_1}{n^2} \sum_{i=1}^{n} \frac{M_i^2(1 - f_{2i})}{m_i} s_{d2i}^2$$

where $\hat{R} = \sum_{i=1}^{n} M_i \bar{y}_{i.} / \sum_{i=1}^{n} M_i \bar{x}_{i.}$ and

$$s^2_{d2i} = \frac{\sum_{j=1}^{m_i} (d_{ij} - \bar{d}_{i.})^2}{m_i - 1} \quad \text{where } d_{ij} = y_{ij} - \hat{R}\, x_{ij}$$

(5.3.23) $\hat{\tau} = M_0\hat{\mu}$, where $\hat{\mu}$ is given by Equation (5.3.21).

C10

(5.3.24) $\hat{V}(\hat{\tau}) = M_0^2\hat{V}(\hat{\mu})$ where $\hat{V}(\hat{\mu})$ is given by Equation (5.3.22).

C10

(5.3.25) $$\hat{\mu} = \mu_x\left(\sum_{i=1}^{n} \bar{y}_{i.} \Big/ \sum_{i=1}^{n} \bar{x}_{i.}\right)$$

(5.3.26) $$\hat{V}(\hat{\mu}) = \frac{1 - f_1}{n} \sum_{i=1}^{n} (\bar{y}_{i.} - \hat{R}\, x_{i.})^2/(n - 1)$$

$$+ \frac{f_1}{n^2} \sum_{i=1}^{n} (1 - f_{2i})\, s^2_{d2i}$$

where $\hat{R} = \bar{y}_{i.} / \sum_{i=1}^{n} \bar{x}_{i.}$ and

$$s^2_{d2i} = \sum_{i=1}^{n} (d_{ij} - \bar{d}_{i.})^2/(m_i - 1) \text{ where } d_{ij} = y_{ij} - \hat{R}\, x_{ij}$$

(5.3.27) $\hat{\tau} = N\, M\, \hat{\mu}$ where $\hat{\mu}$ is given by Equation (5.3.25).

(5.3.28) $\hat{V}(\hat{\tau}) = N^2M^2\hat{V}(\hat{\mu})$ where $\hat{V}(\hat{\mu})$ is given by Equation 5.3.26).

(5.3.29) $$\hat{\mu} = \mu_x\left(\sum_{i=1}^{n} M_i\bar{y}_{i.} \Big/ \sum_{i=1}^{n} M_i\bar{x}_{i.}\right)$$

(5.3.30) $$\hat{V}(\hat{\mu}) = \frac{1 - f_1}{n\bar{M}^2} \sum_{i=1}^{n} \frac{M_i^2(\bar{y}_{i.} - \hat{R}\bar{x}_{i.})^2}{n - 1} + \frac{f_1}{n^2\bar{M}^2} \sum_{i=1}^{n} \frac{M_i^2(1 - f_{2i})}{m}\, s^2_{d2i}$$

where $\hat{R} = \sum_{i=1}^{n} M_i\bar{y}_{i.} / \sum_{i=1}^{n} M_i\bar{x}_{i.}$ and

$$s^2_{d2i} = \sum_{i=1}^{m} (d_{ij} - \bar{d}_{i.})^2/(m - 1) \text{ where } d_{ij} = y_{ij} - \hat{R}\, x_{ij}$$

(5.3.31) $\hat{\tau} = M_0\hat{\mu}$ where $\hat{\mu}$ is given by Equation (5.3.25).

(5.3.32) $\hat{V}(\hat{\tau}) = M_0^2\hat{V}(\hat{\mu})$ where $\hat{V}(\hat{\mu})$ is given by Equation (5.3.26).

Bounds on the estimate of variance

(5.3.33) L1,L6 $\hat{V}_u(\hat{\mu}) = s_{bz}^2/n$ where

$$s_{bz}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\bar{z}_{i.} - \hat{\mu})^2, \text{ where}$$

$$z_{ij} = \frac{M_i}{M_0} N y_{ij} \quad (\text{Note } z_{ij} = y_{ij} \text{ when } M_i = M).$$

(5.3.34) L1,L6 $\hat{V}_L(\hat{\mu}) = (1 - \frac{n}{N})\hat{V}_u(\hat{\mu})$

where $\hat{V}_u(\hat{\mu})$ is as in (5.3.33).

(5.3.35) L1,L6 $\hat{V}_u(\hat{\tau}) = M_0^2\hat{V}_u(\hat{\mu})$ where $\hat{V}_u(\hat{\mu})$ is given by Equation (5.3.33).

(5.3.36) L1,L6 $\hat{V}_L(\hat{\tau}) = (1 - \frac{n}{N})\hat{V}_u(\hat{\tau})$, where $\hat{V}_u(\hat{\tau})$ is as in Equation (9.3.35).

(5.3.37) L7 $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\bar{z}_{i.}$ where

$$z_{ij} = \frac{M_i}{M_0}\frac{y_{ij}}{P_i} \quad (\text{Note } z_{ij} = \frac{y_{ij}}{NP_i} \text{ when } M_i = M).$$

(5.3.38) L7 $\hat{V}(\hat{\mu}) = s_{bz}^2/n$ where $s_{bz}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\bar{z}_{i.} - \hat{\mu})^2$

where z_{ij} is as given in Equation (5.3.37).

(5.3.39) L7 $\hat{\tau} = \frac{M_0}{n}\sum_{i=1}^{n}\bar{z}_{i.}$

(5.3.40) $\hat{V}(\hat{\tau}) = (M_0)^2\hat{V}(\hat{\mu})$ where $\hat{V}(\hat{\mu})$ is as in Equation (5.3.38).

Chapter 6. THREE STAGE SAMPLE DESIGN.

6.1 Introduction: An extension of the two-state sampling design is the three-stage sampling design. It can prove to be very useful in many situations when it is costly or perhaps impossible to measure entire secondary units. We illustrate with examples. Results of the examples are based on the unbiased estimates given later in this chapter.

Example 6.1.1. Given a universe of N=49 townships, seven townships on a side, each township containing M=36 sections, six sections on a side, each section containing B=576 1.1 - acre pixels, 24 pixels on a side, a simple random sample of $n=9$ townships was selected. Within each selected township, a simple random sample of $m=7$ sections was selected, and within each selected section, a simple random sample of b=16 pixels was selected, resulting in a three-stage sample of 1008 pixels to estimate average timber volume per pixel ($\hat{\mu}$ timber volume), average herbage per pixel ($\hat{\mu}$ herbage), and the proportion of pixels in forest area ($\hat{p}$ forest). The results were:

	$\hat{\theta}$	$\sqrt{\hat{V}(\hat{\theta})}$	$100\sqrt{\hat{V}(\hat{\theta})}/\hat{\theta}$
timber volume, $\hat{\theta} = \hat{\mu}$ timber volume	895.8	131.6	14.7%
range forage, $\hat{\theta} = \hat{\mu}$ herbage	1272.3	150.4	11.8%
forest area, $\hat{\theta} = \hat{p}$ forest	0.4514	0.0575	12.7%

Example 6.1.2. Given a universe of N=49 townships, seven townships on a side, each township containing M=2304 9.9-acre blocks, 48 blocks on a side, each block containing B=9 1.1- acre pixels, three pixels on a side, a simple random sample of n=9 townships was selected. Within each selected township, a simple random sample of m = 56 blocks was selected, and within each selected block, a simple random sample of b=2 pixels was selected, resulting in a three-stage sample of 1008 pixels to estimate average timber volume per pixel ($\hat{\mu}$ timber volume), average herbage per pixel ($\hat{\mu}$ herbage), and the proportion of pixels in forest area ($\hat{p}$ forest). The results were:

	$\hat{\theta}$	$\sqrt{\hat{V}(\hat{\theta})}$	$100\sqrt{\hat{V}(\hat{\theta})}/\hat{\theta}$
timber volume, $\hat{\theta} = \hat{\mu}$ timber volume	924.5	137.0	14.8%
herbage, $\hat{\theta} = \hat{\mu}$ herbage	1444.1	152.2	10.5%
forest area, $\hat{\theta} = \hat{p}$ forest	0.4454	0.0630	14.1%

Example 6.1.3. Given a universe of N=1764 sections, 42 sections on a side, each section containing M=64 9.9 - acre blocks, eight blocks on a side, each block containing B=9 1.1 - acre pixels, three pixels on a side, a simple random sample of n=72 sections was selected. Within each selected section, a simple random sample of m=7 blocks was selected, and within each selected block, a simple random sample of b=2 pixels was selected, resulting in a three-stage sample of 1008 pixels to estimate average timber volume per pixel ($\hat{\mu}$ timber volume), average herbage per pixel ($\hat{\mu}$ herbage), and the proportion of pixels in forest area ($\hat{p}$ forest). The results were:

	$\hat{\theta}$	$\sqrt{\hat{V}(\hat{\theta})}$	$100\sqrt{\hat{V}(\hat{\theta})}/\hat{\theta}$
timber volume, $\hat{\theta} = \hat{\mu}$ timber volume	1115.2	83.6	7.5%
herbage, $\hat{\theta} = \hat{\mu}$ herbage	1127.3	85.3	7.6%
forest area, $\hat{\theta} = \hat{p}$ timber	0.5159	0.0328	6.4%

The formal definition of a three-stage sampling design is given below.

Definition 6.1.1. Three-Stage Sampling Design. Consider a universe that is divided into N groups so that each item in the universe is in exactly one group. These N groups are called first-stage (primary) sampling units. Each first-stage unit is divided into sub-units (called second-stage units) so that the ith primary unit contains M_i second-stage units and each item in the ith primary unit is contained in exactly one second-stage unit. Each second-stage unit is divided into sub-units (called third-stage units); if the ijth second- stage unit contains B_{ij} third-stage units and if each item in the ijth second-stage unit is contained in exactly one third-stage unit the resulting design is called a three-stage sampling design.

Note. The definition does not state how the first, second, or third-stage units are sampled. For example n first-stage units may be sampled from the N universe units by simple random sampling. Then from the ith sampled first-stage unit m_i of M_i second-stage units may be sampled by systematic sampling or by simple random sampling. Finally b_{ij} third-stage units may be selected from the B_{ij} third-stage units in the jth second-stage unit of the ith first-stage unit. The b_{ij} third-stage units from the ijth third-stage unit could be selected by simple random sampling, systematic sampling, or other methods of sampling.

Two types of estimates are considered--unbiased and ratio. In Table 6.1.1 is exhibited the population and sample sizes of each stage, the method of selecting the sample, the type of estimate, and the equation number for the estimate of μ, τ, and p. The formula for sample size is also referenced.

Table 6.1.1

Unit Size / Sample Size		Method of Selection			Type of 2/ Estimate	Equation #			
						TOTAL		MEAN AND PROPORTION	
1st Stage	2nd Stage	1st Stage	2nd Stage	3rd Stage		Point Estimate	Variance Estimate	Point Estimate	Variance Estimate
M_i/m_i	B_{ij}/b_{ij}	SRS	SRS	SRS	UU	6.1.1	6.1.2	6.1.3	6.1.4
M/m	B/b					6.1.5	6.1.6	6.1.7	6.1.8
M_i/m_i	B_{ij}/b	SRS	SRS	SRS	UR	6.1.9	1/	6.1.10	1/
M/m	B/b					6.1.11		6.1.12	
M_i/m_i	B_{ij}/b_{ij}	SRS	SRS	SRS	RU	6.1.13	1/	6.1.14	1/
M/m	B/b					6.1.15		6.1.16	
M_i/m_i	B_{ij}/b_{ij}	SRS	SRS	SRS	RR	6.1.17	1/	6.1.18	1/
M/m	B/b					6.1.19		6.1.20	
M_i/m_i	B_{ij}/b_{ij}	SRS	SRS	SYS	UU	6.1.1	6.1.21	6.1.3	6.1.23
							6.1.22		6.1.26
M/m	B/b	SRS	SRS	SYS	UU	6.1.5	6.1.24	6.1.7	6.1.25
							6.1.27		6.1.28
M_i/m_i	B_{ij}/b_{ij}	SRS	SYS	SRS or SYS	UU	6.1.1	6.1.29	6.1.3	6.1.31
							6.1.30		6.1.32
M/m	B/b	SRS	SYS	SRS or SYS	UU	6.1.5	6.1.33	6.1.7	6.1.35
							6.1.34		6.1.36
M_i/m_i	B_{ij}/b_{ij}	SYS	SRS or SYS	SRS or SYS	UU	6.1.1	1/	6.1.3	1/
M/m	B/b	SYS	SRS or SYS	SRS or SYS	UU	6.1.5	1/	6.1.7	1/
M_i/m_i	B_{ij}/b_{ij}	Random Sample with replacement	SRS or SYS	SRS or SYS	UU	6.1.1	6.1.37	6.1.3	6.1.38
M/m	B/b	Random Sample with replacement	SRS or SYS	SRS or SYS	UU	6.1.5	6.1.39	6.1.7	6.1.40

1/ Generally accepted methods not available without making further assumptions.

2/ The symbols U and R denote respectively unbiased and ratio estimators. For example UR means to estimate population total of the first stage units a ratio estimate is used. This symbol is given by $Y_{i\cdot}^{*R}$ in equation 6.1.9). Then to estimate the population total, an unbiased estimate is used. The symbol is given by $Y_{\cdot\cdot}^{UR}$ (a U in the first position) in equation 6.1.9).

The notaton used for the three-stage sampling design will be the same as the notation for the two-stage design, with the following additions:

B_{ij} = the number of third-stage units in the jth second-stage unit of the ith first-stage unit.

($j = i, 2, \ldots, M_i\, j$; $i = 1, 2, \ldots, N$)

Y_{ijk} = the value of the kth third stage unit of the jth second-stage unit of the ith first-stage unit.

($k = 1, 2, \ldots, B_{ij}$; $j = 1, 2, \ldots, M_i$; $i = 1, 2, \ldots, N$)

b_{ij} = the number of third-stage units sampled in the jth sampled second-stage unit of the ith sampled first-stage unit.

($j = 1, 2, \ldots, m_i$; $i = 1, 2, \ldots, n$)

y_{ijk} = the value of the kth sampled third-stage unit of the jth sampled second-stage unit of the ith sampled first-stage unit.

($k = 1, 2, \ldots, b_{ij}$; $j = 1, 2, \ldots, m_i$; $i = 1, 2, \ldots n$)

$$B.. = \sum_{i=1}^{N} \sum_{j=1}^{M_i} B_{ij}$$

= total number of third-stage units in the population

$Y...$ = population total.

NOTE: For the case of equal sample sizes, M_i, B_{ij}, m_i, b_{ij} are replaced with M, B, m, b, respectively.

Notation

$$s^2(y_{i_1, i_2, \ldots, i_{k-1}, i_k} | i_k, t)$$

$$= \frac{\sum_{i_k=1}^{t} (y_{i_1, i_2 \ldots, i_k} - \bar{y}_{i_1, i_2, \ldots, i_{k-1}.})^2}{t - 1}$$

where

$$\bar{y}_{i_1, i_2, \ldots, i_{k-1}, .} = t^{-1} \sum_{i_k=1}^{t} y_{i_1, i_2, \ldots, i_k}.$$

(6.1.1) $$\hat{\tau} = \hat{Y}^{UU}_{\ldots} = \frac{N}{n} \sum_{i=1}^{n} \hat{Y}^{*U}_{i..}$$

where $$\hat{Y}^{*U}_{i..} = \frac{M_i}{m_i} \sum_{j=1}^{m_i} \hat{Y}_{ij.}$$

where $$\hat{Y}_{ij.} = \frac{B_{ij}}{b_{ij}} \sum_{k=1}^{b_{ij}} y_{ijk}$$

(6.1.2) $$\hat{V}(\hat{Y}^{UU}_{\ldots}) = \frac{N(N-n)}{n} s^2(\hat{Y}^{*U}_{i..} | i,n) + \frac{N}{n} \sum_{i=i}^{n} \hat{V}(\hat{Y}^{*U}_{i..})$$

where $$\hat{V}(\hat{Y}^{*U}_{i..}) = \frac{M_i(M_i - m_i)}{m_i} s^2(y_{ij.} | j, m_i)$$

$$+ \frac{M_i}{m_i} \sum_{j=1}^{m_i} \hat{V}(\hat{Y}_{ij.})$$

where $$y_{ij.} = \sum_{k=1}^{b_{ij}} y_{ijk}$$

and where $\hat{V}(\hat{Y}_{ij.}) = \dfrac{B_{ij}(B_{ij}-b_{ij})}{b_{ij}} s^2(y_{ijk} \mid k, b_{ij})$

(6.1.3) $\underset{L1,L2,L4,L5}{\hat{\mu}} = \hat{\tau} \Big/ \sum_{i=1}^{N} \sum_{j=1}^{M_i} B_{ij}$, where $\hat{\tau}$ is given in Equation (6.1.1).

(6.1.4) $\underset{L1,L2,L4,L5}{\hat{V}(\hat{\mu})} = \hat{V}(\hat{\tau})/(\sum_{i=1}^{N} \sum_{j=1}^{M_i} B_{ij})^2$, where $\hat{V}(\hat{\tau})$ is given in Equation (6.2.2).

(6.1.5) $\hat{\tau} = \hat{Y}^{UU}_{...} = \dfrac{N}{n} \sum_{i=1}^{n} \hat{Y}^{*U}_{i..}$ where $\hat{Y}^{*U}_{i..} = \dfrac{M}{m} \sum_{j=1}^{m} \hat{Y}_{ij.}$

where $\hat{Y}_{ij.} = \dfrac{B}{b} \sum_{k=1}^{b} y_{ijk}$

(6.1.6) $\hat{V}(\hat{Y}^{UU}_{...}) = \dfrac{N(N-n)}{n} s^2(\hat{Y}^{*U}_{i..} \mid i,n) + \dfrac{N}{n} \sum_{i=1}^{n} \hat{V}(\hat{Y}^{*U}_{i..})$

where $\hat{V}(\hat{Y}^{*U}_{i..}) = \dfrac{M(M-m)}{m} s^2(Y_{ij.} \mid j,m) + \dfrac{M}{m} \sum_{j=1}^{m} \hat{V}(\hat{Y}_{ij.})$

where $\hat{V}(\hat{Y}_{ij.}) = \dfrac{B(B-b)}{b} s^2(y_{ijk} \mid k,b)$

(6.1.7) $\hat{\mu} = \hat{\tau}/NMB$, where $\hat{\tau}$ is given in Equation (6.1.5).

(5.1.8) $\hat{V}(\hat{\mu}) = \hat{V}(\hat{\tau})/(NMB)^2$, where $\hat{V}(\hat{\mu})$ is given in Equation (6.1.6).

(6.1.9) $\hat{\tau} = \hat{Y}^{UR}_{...} = \dfrac{N}{n} \sum_{i=1}^{n} \hat{Y}^{*R}_{i..}$ where $\hat{Y}^{*R}_{i..} = X_{i..} \dfrac{\sum_{j=1}^{m_i} B_{ij} \bar{y}_{ij.}}{\sum_{j=1}^{m_i} B_{ij} \bar{x}_{ij}}$

(6.1.10) $\hat{\mu} = \hat{\tau} / \sum_{i=1}^{N} \sum_{j=1}^{M} B_{ij}$ where $\hat{\tau}$ is given by Equation (6.1.9).

(6.1.11) $\hat{\tau} = \hat{Y}^{UR}_{...} = \dfrac{N}{n} \sum_{i=1}^{n} \hat{Y}^{*R}_{i..}$ where $\hat{Y}^{*R}_{i..} = X_{i..} \dfrac{\sum_{j=1}^{m} \bar{y}_{ij}}{\sum_{j=1}^{m} \bar{x}_{ij}}$

(6.1.12) $\hat{\mu} = \hat{\tau}/NMB$, where $\hat{\tau}$ is given by Equation (6.1.11).

(6.1.13) $\hat{\tau} = \hat{Y}^{RU}_{...} = X_{...} \dfrac{\sum_{i=1}^{n} Y^{*U}_{i..}}{\sum_{i=1}^{n} X_{i..}}$ where $\hat{Y}^{*U}_{i..}$ is given by Equation (6.1.1).

(6.1.14) $\hat{\mu} = \hat{\tau}/\sum_{i=1}^{N}\sum_{j=1}^{M_i} B_{ij}$ where $\hat{\tau}$ is given by Equation (6.1.13)

(6.1.15) $\hat{\tau} = \hat{Y}^{RU}_{...} = X_{...} \dfrac{\sum_{i=1}^{n} Y^{*U}_{i..}}{\sum_{i=1}^{n} X_{i..}}$ where $\hat{Y}^{*U}_{i..}$ is given by Equation (6.1.5)

(6.1.16) $\hat{\mu} = \hat{\tau}/NMB$ where $\hat{\tau}$ is given by Equation (6.1.15)

(6.1.17) $\hat{\tau} = \hat{Y}^{RR}_{...} = X_{...} \dfrac{\sum_{i=1}^{n} Y^{*R}_{i..}}{\sum_{i=1}^{n} X_{i..}}$ where $\hat{Y}^{*R}_{i..}$ is given by Equation (6.1.9)

(6.1.18) $\hat{\mu} = \hat{\tau}/\sum_{i=1}^{N}\sum_{j=1}^{M_i} B_{ij}$.

(6.1.19) $\hat{\tau} = \hat{Y}^{RR}_{...} = X_{...} \dfrac{\sum_{i=1}^{n} \hat{Y}^{*R}_{i..}}{\sum_{i=1}^{n} X_{i..}}$ where $\hat{Y}^{*R}_{i..}$ is given by Equation (6.1.11)

(6.1.20) $\hat{\mu} = \hat{\tau}/NMB$ where $\hat{\tau}$ is given by Equation (6.1.19)

Bounds on $\hat{V}(\hat{\tau})$ are given by Equations (6.1.21) and (6.1.22).

(6.1.21) L1,L5,L6 $\hat{V}_U(\hat{Y}^{UU}_{...}) = \dfrac{N(N-n)}{n} s^2(\hat{Y}^{*U}_{i..} | i, n) + \dfrac{N}{n}\sum_{i=1}^{n} \hat{V}_U(\hat{Y}^{*U}_{i..})$

where $\hat{V}_U(\hat{Y}^{*U}_{i..}) = \dfrac{M_i^2}{m_i} s^2(\hat{Y}_{ij.} | j, m_i)$

(6.1.22) L1,L5,L6
$$\hat{V}_L(\hat{Y}^{UU}_{...}) = \frac{N(N-n)}{n} s^2(\hat{Y}^{*U}_{i..} \mid i,k) + \frac{N}{n}\sum_{i=1}^{n} \hat{V}_L(\hat{Y}^{*U}_{i..})$$

where

$$\hat{V}_L(\hat{Y}^{*U}_{i..}) = \frac{M_i(M_i-m_i)}{m_i} s^2(\hat{Y}_{ij.} \mid j,m_i).$$

(6.1.23) L1,L5,L6
$$\hat{V}_L(\hat{\mu}) = \hat{V}_L(\hat{Y}^{UU}_{...})/(\sum_{i=1}^{N}\sum_{j=1}^{M_i} B_{ij})^2 \quad \text{where}$$

$\hat{V}_L(\hat{Y}^{UU}_{...})$ is given by Equation (6.1.22).

(6.1.24) L1,L5,L6
$$\hat{V}_L(\hat{Y}^{UU}_{...}) = \frac{N(N-n)}{n} s^2(\hat{Y}^{*U}_{i..} \mid i,n) + \frac{N}{n}\sum_{i=1}^{n} \hat{V}_L(\hat{Y}^{*U}_{i..})$$

where $\hat{V}_L(\hat{Y}^{*U}_{i..}) = \frac{M(M-m)}{m} s^2(\hat{Y}_{ij.} \mid j,m)$

(6.1.25) L1,L5,L6
$$\hat{V}_L(\hat{\mu}) = \hat{V}_L(\hat{Y}^{UU}_{...})/(NMB)^2 \quad \text{where}$$

$\hat{V}_L(\hat{Y}^{UU}_{...})$ is given by Equation(6.1.24).

Bounds on $\hat{V}(\hat{\mu})$ are given by Equation (6.1.23) and

(6.1.26) $\hat{V}_U(\hat{\mu}) = \hat{V}_U(\hat{Y}^{UU}_{...})/\sum_{i=1}^{N}\sum_{j=1}^{M_i} B_{ij})^2$, where $\hat{V}_U(\hat{Y}^{UU}_{...})$ is given by Equation (6.1.21

Bounds on $\hat{V}(\hat{\tau})$ are given by Equation (6.1.24) and

(6.1.27)
$$\hat{V}_U(\hat{Y}^{UU}_{...}) = \frac{N-n}{n} s^2(\hat{Y}^{*U}_{i..} \mid i,n) + \frac{N}{n}\sum_{i=1}^{n} \hat{V}_u(\hat{Y}^{*U}_{i..})$$

where $\hat{V}_U(\hat{Y}^{*U}_{i..}) = \frac{M^2}{m} s^2(\hat{Y}_{ij.}|j,m)$.

Bounds on $\hat{V}(\hat{\tau})$ are given by Equation (6.1.25) and

(6.1.28) $\hat{V}_U(\hat{\mu}) = \hat{V}_U(\hat{Y}^{UU}_{...})/(NMB)^2$ where $\hat{V}_U(\hat{Y}^{UU}_{...})$ is given by Equation (6.1.27).

Bounds on $\hat{V}(\hat{\tau})$ are given by

(6.1.29) $\hat{V}_L(\hat{Y}^{UU}_{...}) = \frac{N(N-n)}{n} s^2(\hat{Y}^{*U}_{i..}|i,n)$
L1,L5,L6

and

(6.1.30) $\hat{V}_U(\hat{Y}^{UU}_{...}) = \frac{N^2}{n} s^2(\hat{Y}^{*U}_{i..}|i,n)$ where $\hat{Y}^{*U}_{i..}$ is given by Equation (6.1.1).
L1,L5,L6

Bounds on $\hat{V}(\hat{\mu})$ are given by

(6.1.31) $\hat{V}_L(\hat{\mu}) = \hat{V}_L(\hat{Y}^{UU}_{...})/(\sum_{i=1}^{N}\sum_{j=1}^{M_i} B_{ij})^2$ where $\hat{V}_L(\hat{Y}^{UU}_{...})$ is given by
L1,L5,L6

Equation (6.1.29)

and

(6.1.32) $\hat{V}_U(\hat{\mu}) = \hat{V}_U(\hat{Y}^{UU}_{...})/(\sum_{i=1}^{N}\sum_{j=1}^{M_i} B_{ij})^2$ where $\hat{V}_U(\hat{Y}^{UU}_{...})$ is given by Equation (6.1.30).
L1,L5,L6

Bounds on $\hat{V}(\hat{\tau})$ are given by

(6.1.33) $\hat{V}_L(\hat{Y}^{UU}_{...}) = \frac{N(N-n)}{n} s^2(\hat{Y}^{*U}_{i..}|i,n)$

and

(6.1.34) $\hat{V}_U(\hat{Y}^{UU}_{...}) = \frac{N^2}{n} s^2(\hat{Y}^{*U}_{i..}|i,n)$ where $\hat{Y}^{*U}_{i..}$ is given by Equation (6.1.5).

Bounds on $\hat{V}(\hat{\mu})$ are given by

(6.1.35) $\hat{V}_L(\hat{\mu}) = \hat{V}_L(\hat{Y}^{UU}_{...})/(NMB)^2$ where $\hat{V}_L(\hat{Y}^{UU}_{...})$ is given by Equation (6.1.33).

(6.1.36) $\hat{V}_U(\hat{\mu}) = \hat{V}_U(\hat{Y}^{UU}_{...})/(NMB)^2$ where $\hat{V}_U(\hat{Y}^{UU}_{...})$ is given by Equation (6.1.34).

(6.1.37) $\hat{V}(\hat{\tau}) = \frac{N^2}{n} s^2(\hat{Y}^{*U}_{i..}|i,n)$ where $\hat{Y}^{*U}_{i..}$ is given by Equation (6.1.1).

(6.1.38) $\hat{V}(\hat{\mu}) = \hat{V}(\hat{\tau})/(\sum_{i=1}^{N}\sum_{j=1}^{M_i} B_{ij})^2$, where $\hat{V}(\hat{\tau})$ is given by Equation (6.1.37)

(6.1.39) $\hat{V}(\hat{\tau}) = \frac{N^2}{n} s^2_{(\hat{Y}^{*U}_{i..}|i,n)}$ where $\hat{Y}^{*U}_{i..}$ is given by Equation (6.1.5).

(6.1.40) $\hat{V}(\hat{\mu}) = \hat{V}(\hat{\tau})/(NMB)^2$, where $\hat{V}(\hat{\tau})$ is given by Equation (6.1.39).

Chapter 7. FOUR-STAGE SAMPLING.

7.1 Four-stage sampling is the next extension of multi-stage sampling. As with the extension of two-stage sampling to three-stage sampling, the complexity of estimators and possibilities of sample selection procedures are increased. Four-stage sampling is the last multi-stage design covered in this monograph. An example of four-stage sampling is now presented.

Example 7.1.1. Given a universe of N=49 townships, seven townships on a side, each township containing M=36 sections, six sections on a side, each section containing B=64 9.9 - acre blocks, eight blocks on a side, each block containing C=9 1.1 - acre pixels, three pixels on a side, a simple random sample of n=7 townships was selected. Within each selected township, a simple random sample of m=6 sections was selected. Within each selected section, a simple random sample of b=8 blocks was selected, and within each selected block, a simple random sample of c=3 pixels was selected, resulting in a four-stage sample of 1008 pixels to estimate average timber volume per pixel ($\hat{\mu}$ timber volume), average herbage per pixel ($\hat{\mu}$ herbage), and the proportion of forest area ($\hat{p}$ forest). The results were:

	$\hat{\theta}$	$\sqrt{\hat{V}(\hat{\theta})}$	$100\sqrt{\hat{V}(\hat{\theta})}/\hat{\theta}$
timber volume, $\hat{\theta} = \hat{\mu}$ timber volume	809.3	147.7	18.3%
herbage, $\hat{\theta} = \hat{\mu}$ herbage	1476.8	173.0	11.7%
forest area, $\hat{\theta} = \hat{p}$ forest	0.3978	0.0668	16.8%

The formal definition of four-stage sampling is given below.

Definition 7.1.1. Four- Stage Sampling Design. Consider a universe that is divided into N groups so that each item in the universe is in exactly one group. These N groups are called first-stage (primary) sampling units. Each first-stage unit is divided into sub-units (called second-stage units) so that the ith primary unit contains M_i second-stage units and each item in the ith primary unit is contained in exactly one second-stage unit.

Each second-stage unit is divided into sub-units (called third-stage units) so that the ijth second-stage unit contains B_{ij} third-stage units and each item in the ijth second-stage unit is contained in exactly one third-stage unit. Each third-stage unit is divided into sub-units (called fourth-stage units); if the ijkth third-stage unit contains C_{ijk} fourth-stage units and if each item in the ijkth third-stage unit is contained in exactly one fourth-stage unit the resulting design is called a four-stage sampling design.

Note. The definition does not state how the first, second, third, or fourth-stage units are sampled. Different combinations of simple random sampling and systematic sampling are available.

Various combinations of simple random sampling and systematic sampling are considered as well as two types of estimates--unbiased and ratio.

The notation for the four-stage sampling design will be the same as the notation for three-stage sampling, with the following additions.

C_{ijk} = the number of fourth-stage units in the k-th third-stage unit of the j-th second-stage unit of the i-th first-stage unit.

($k = 1, 2, \ldots, B_{ij}$; $j = 1, 2, \ldots, M_i$; $i = 1, 2, \ldots, N$)

$Y_{ijk\ell}$ = the value of the ℓ-th fourth-stage unit of the (i,j,k)-th third-stage unit. ($\ell = 1, 2, \ldots, C_{ijk}$; $k = 1, 2, \ldots, B_{ij}$; $j = 1, 2, \ldots, M_i$; $i = 1, 2, \ldots, N$)

c_{ijk} = the number of sampled fourth-stage units in the k-th sampled third-stage unit of the j-th sampled second-stage unit of the i-th sampled first-stage unit.

($k = 1, 2, \ldots, b_{ij}$; $j = 1, 2, \ldots, m_i$; $i = 1, 2, \ldots, n$)

$y_{ijk\ell}$ = the value of the ℓ-th sampled fourth-stage unit of the (i,j,k) -th sampled third-stage unit.

($\ell = 1, 2, \ldots, c_{ijk}$; $k = 1, 2, \ldots, b_{ij}$; $j=1, 2, \ldots, m_i$: $i = 1,2, \ldots, n$)

Note: For the case of equal sample sizes, M_i, B_{ij}, C_{ijk}, m_i, b_{ij}, c_{ijk} are replaced with M,B,C, m,b,c, respectively.

$$C... = \sum_{i=1}^{N} \sum_{j=1}^{M_i} \sum_{k=1}^{B_{ij}} C_{ijk}$$

= total number of fourth-stage units in the population.

$Y.... =$ Population total.

Equation numbers for unbiased and ratio estimates of μ, τ and p are given in Table 7.1.1 for unequal and equal sample sizes for various sample selection procedures.

The third-state unit totals $\hat{Y}_{ijk.}$ are estimated using the unbiased estimator. We can replace it by the following ratio estimator, wherever $\hat{Y}_{ijk.}$ is used.

$$Y_{ijk.}^{***R} = X_{ijk.} \frac{\sum_{\ell=1}^{C_{ijk}} Y_{ijk\ell}}{\sum_{\ell=1}^{C_{ijk}} X_{ijk\ell}}$$

Notice that

1) When we use $Y_{ijk.}^{***R}$ we do not have an estimate of variance.

2) $C_{ijk} = C$, for the case of equal cluster and sample sizes.

A similar ratio estimator can be used for three-stage sampling.

Note: Whenever the above is done, the cluster totals cannot replace the auxiliary variables.

Table 7.1.1

Unit Size: / Sample Size	Method of Selection: 1st Stage	2nd Stage	3rd Stage	4th Stage	Type of Estimate [2/]	Equation #: TOTAL Point Estimate	TOTAL Variance Estimate	MEAN AND PROPORTION Point Estimate	MEAN AND PROPORTION Variance Estimate
Unequal Equal	SRS	SRS	SRS	SRS	RRR	7.1.1 7.1.3	1/	7.1.2 7.1.4	1/
Unequal Equal	SRS	SRS	SRS	SRS	URR	7.1.5 7.1.7	1/	7.1.6 7.1.8	1/
Unequal Equal	SRS	SRS	SRS	SRS	UUR	7.1.9 7.1.11	1/	7.1.10 7.1.12	1/
Unequal Equal	SRS	SRS	SRS	SRS	RUR	7.1.13 7.1.15	1/	7.1.14 7.1.16	1/
Unequal Equal	SRS	SRS	SRS	SRS	RUU	7.1.17 7.1.19	1/	7.1.18 7.1.20	1/
Unequal Equal	SRS	SRS	SRS	SRS	RRU	7.1.21 7.1.23	1/	7.1.22 7.1.24	1/
Unequal Equal	SRS	SRS	SRS	SRS	URU	7.1.25 7.1.27	1/	7.1.26 7.1.28	1/
Unequal Equal	SRS	SRS	SRS	SRS	UUU	7.1.29 7.1.33	7.1.30 7.1.34	7.1.31 7.1.35	7.1.32 7.1.36
Unequal Equal	SRS	SRS	SRS	SYS	UUU	7.1.29 7.1.33	7.1.37, 7.1.38 7.1.41, 7.1.42	7.1.31 7.1.35	7.1.39, 7.1.40 7.1.43, 7.1.44
Unequal Equal	SRS	SRS	SYS	SRS or SYS	UUU	7.1.29 7.1.33	7.1.45, 7.1.46 7.1.49, 7.1.50	7.1.31 7.1.35	7.1.47, 7.1.48 7.1.51, 7.1.52
Unequal Equal	SRS	SYS	SRS or SYS	SRS or SYS	UUU	7.1.29 7.1.33	7.1.53, 7.1.54 7.1.57, 7.1.58	7.1.31 7.1.35	7.1.55, 7.1.56 7.1.59, 7.1.60
Unequal Equal	SYS	SRS or SYS	SRS or SYS	SRS or SYS	UUU	7.1.29 7.1.33	1/	7.1.31 7.1.35	1/
Unequal Equal	Random Sample with Replacement	SRS or SYS	SRS or SYS	SRS or SYS	UUU	7.1.29 7.1.33	7.1.54 7.1.58	7.1.31 7.1.35	7.1.56 7.1.60

1/ Not generally available without additional assumptions.

2/ The symbols U and R denote respectively unbiased and ratio estimators.

(7.1.1) $$\hat{\tau} = \hat{Y}^{RRR}_{....} = X.... \frac{\sum_{i=1}^{n} \hat{Y}^{*RR}_{i..}}{\sum_{i=1}^{n} X_{i}...} \text{ where } \hat{Y}^{*RR}_{i..} = X_{i}... \frac{\sum_{j=1}^{m_i} \hat{Y}^{**R}_{ij..}}{\sum_{j=1}^{m_i} X_{ij..}}$$

where $$\hat{Y}^{**R}_{ij..} = X_{ij..} \frac{\sum_{k=1}^{b_{ij}} \hat{Y}_{ijk.}}{\sum_{k=1}^{b_{ij}} X_{ijk.}}$$

where $$\hat{Y}_{ijk.} = \frac{C_{ijk}}{c_{ijk}} \sum_{\ell=1}^{c_{ijk}} y_{ijk\ell}$$

(7.1.2) $\hat{\mu} = \hat{\tau}/C...$ where $\hat{\tau}$ is defined by Equation (7.1.1).

(7.1.3) $$\hat{\tau} = \hat{Y}^{RRR}_{....} = X.... \frac{\sum_{i=1}^{n} \hat{Y}^{*RR}_{i...}}{\sum_{i=1}^{n} X_{i}...} \text{ where } \hat{Y}^{*RR}_{i...} = X_{i}... \frac{\sum_{j=1}^{m} \hat{Y}^{**R}_{ij..}}{\sum_{j=1}^{m} X_{ij..}}$$

where $$\hat{Y}^{**R}_{ij..} = X_{ij..} \frac{\sum_{k=1}^{b} \hat{Y}_{ijk.}}{\sum_{k=1}^{b} X_{ijk.}} \text{ where } \hat{Y}_{ijk.} = \frac{C}{c} \sum_{\ell=1}^{c} y_{ijk\ell}$$

(7.1.4) $\hat{\mu} = \hat{\tau}/(NMBC)$ where $\hat{\tau}$ is defined by Equation (7.1.3).

(7.1.5) $$\hat{\tau} = \hat{Y}^{URR}_{....} = \frac{N}{n} \sum_{i=1}^{n} \hat{Y}^{*RR}_{i...}$$ where $\hat{Y}^{*RR}_{i...}$ is given in Equation (7.1.1).

(7.1.6) $\hat{\mu} = \hat{\tau}/C...$, where $\hat{\tau}$ is defined by Equation (7.1.5).

(7.1.7) $$\hat{\tau} = \hat{Y}^{URR}_{....} = \frac{N}{n} \sum_{i=1}^{n} \hat{Y}^{*RR}_{i...}$$ where $\hat{Y}^{*RR}_{i...}$ is given in Equation (7.1.3).

(7.1.8) $\hat{\mu} = \hat{\tau}/(NMBC)$ where $\hat{\tau}$ is defined by Equation (7.1.7).

(7.1.9) $\hat{\tau} = \hat{Y}^{UUR}_{....} = \frac{N}{n} \sum_{i=1}^{n} \hat{Y}^{*UR}_{i...}$ where $\hat{Y}^{*UR}_{i...} = \frac{M_i}{m_i} \sum_{j=1}^{m_i} \hat{Y}^{**R}_{ij..}$ where

$\hat{Y}^{**R}_{ij..}$ is given in Equation (7.1.1).

(7.1.10) $\hat{\mu} = \hat{\tau}/C...$, where $\hat{\tau}$ is defined by Equation (7.1.9).

(7.1.11) $\hat{\tau} = \hat{Y}^{UUR}_{....} = \frac{N}{n} \sum_{i=1}^{n} \hat{Y}^{*UR}_{i...}$ where $\hat{Y}^{**R}_{i...} = \frac{M}{m} \sum_{j=1}^{m} \hat{Y}^{**R}_{ij..}$ where $\hat{Y}^{**R}_{ij..}$
is given in Equation (7.1.3).

(7.1.12) $\hat{\mu} = \hat{\tau}/(NMBC)$ where $\hat{\tau}$ is defined by Equation (7.1.11).

(7.1.13) $\hat{\tau} = \hat{Y}^{RUR}_{....} = X.... \dfrac{\sum_{i=1}^{n} \hat{Y}^{*UR}_{i...}}{\sum_{i=1}^{n} X_{i...}}$ where $\hat{Y}^{*UR}_{i...}$ is given in Equation (7.1.9).

(7.1.14) $\hat{\mu} = \hat{\tau}/C...$ where $\hat{\tau}$ is defined by Equation (7.1.13).

(7.1.15) $\hat{\tau} = \hat{Y}^{RUR}_{....} = X.... \dfrac{\sum_{i=1}^{n} \hat{Y}^{*UR}_{i...}}{\sum_{i=1}^{n} X_{i...}}$ where $\hat{Y}^{*UR}_{i...}$ is given in Equation (7.1.11).

(7.1.16) $\hat{\mu} = \hat{\tau}/(NMBC)$ where $\hat{\tau}$ is defined by Equation (7.1.15).

(7.1.17) $\hat{\tau} = \hat{Y}^{RUU}_{....} = X.... \dfrac{\sum_{i=1}^{n} \hat{Y}^{*UU}}{\sum_{i=1}^{n} X_{i...}}$ where $\hat{Y}^{*UU}_{i...} = \frac{M_i}{m_i} \sum_{j=1}^{m_i} \hat{Y}^{**U}_{ij..}$

where $\hat{Y}^{**U}_{ij..} = \frac{B_{ij}}{b_{ij}} \sum_{k=1}^{b_{ij}} \hat{Y}_{ijk.}$ where $\hat{Y}_{ijk.}$ is given in Equation (7.1.1).

(7.1.18) $\hat{\mu} = \hat{\tau}/C...$ where $\hat{\tau}$ is defined by Equation (7.1.17).

(7.1.19) $\hat{\tau} = \hat{Y}^{RUU}_{....} = X.... \dfrac{\sum_{i=1}^{n} \hat{Y}^{*UU}}{\sum_{i=1}^{n} X_{i...}}$ where $\hat{Y}^{*UU}_{i...} = \dfrac{M}{m} \sum_{j=1}^{m} \hat{Y}^{**U}_{ij..}$

where $\hat{Y}^{**U}_{ij..} = \dfrac{B}{b} \sum_{k=1}^{b} \hat{Y}_{ijk.}$ where $\hat{Y}_{ijk.}$ is given in Equation (7.1.3).

(7.1.20) $\hat{\mu} = \hat{\tau}/(NMBC)$ where $\hat{\tau}$ is defined by Equation (7.1.19).

(7.1.21) $\hat{\tau} = \hat{Y}^{RRU}_{....} = X.... \dfrac{\sum_{i=1}^{n} \hat{Y}^{*RU}_{i...}}{\sum_{i=1}^{n} X_{i...}}$ where $\hat{Y}^{*RU}_{i...} = X_{i...} \dfrac{\sum_{j=1}^{m_i} \hat{Y}^{**U}_{ij..}}{\sum_{j=1}^{m_i} X_{ij..}}$

where $\hat{Y}^{**U}_{ij..}$ is given in Equation (7.1.17).

(7.1.22) $\hat{\mu} = \hat{\tau}/C...$ where $\hat{\tau}$ is defined by Equation (7.1.21).

(7.1.23) $\hat{\tau} = \hat{Y}^{RRU}_{....} = X.... \sum_{i=1}^{n} \hat{Y}^{*RU}_{i...}$ where $\hat{Y}^{*RU}_{i...} = X_{i...} \dfrac{\sum_{j=1}^{m} \hat{Y}^{**U}_{ij..}}{\sum_{j=1}^{m} X_{ij..}}$

where $\hat{Y}^{**U}_{ij..}$ is given in Equation (7.1.19).

(7.1.24) $\hat{\mu} = \hat{\tau}/(NMBC)$ where $\hat{\tau}$ is defined by Equation (7.1.23).

(7.1.25) $\hat{\tau} = \hat{Y}^{URU}_{....} = \dfrac{N}{n} \sum_{i=1}^{n} \hat{Y}^{*RU}_{i...}$ where $\hat{Y}^{*RU}_{i...}$ is given in Equation (7.1.21).

(7.1.26) $\hat{\mu} = \hat{\tau}/C...$ where $\hat{\tau}$ is defined by Equation (7.1.25).

(7.1.27) $\hat{\tau} = \hat{Y}^{URU}_{....} = \dfrac{N}{n} \sum_{i=1}^{n} \hat{Y}^{*RU}_{i...}$ where $\hat{Y}^{*RU}_{i...}$ is given in Equation (7.1.23).

(7.1.28) $\hat{\mu} = \hat{\tau}/(NMBC)$ where $\hat{\tau}$ is defined by Equation (7.1.28).
L1,L2,L4,L5

(7.1.29) $\hat{\tau} = \hat{Y}^{UUU}_{....} = \dfrac{N}{n} \sum_{i=1}^{n} \hat{Y}^{*UU}_{i...}$ where $\hat{Y}^{*UU}_{i...}$ is given in Equation (7.1.17).
L1,L2,L4,L5

(7.1.30) $\hat{V}(\hat{\tau}) = \frac{N(N-n)}{n} s^2_{(\hat{Y}^{*UU}_{i...}|i,n)} + \frac{N}{n} \sum_{i=1}^{n} \hat{V}(\hat{Y}^{*UU}_{i...})$ where $\hat{Y}^{*UU}_{i...}$

is given in Equation (7.1.17) and

$$\hat{V}(\hat{Y}^{*UU}_{i...}) = \frac{M_i(M_i-m_i)}{m_i} s^2_{(\hat{Y}^{**U}_{ij..}|j,m_i)} + \frac{M_i}{m_i} \sum_{j=1}^{m_i} \hat{V}(\hat{Y}^{**U}_{ij..})$$

where $\hat{Y}^{**U}_{ij..}$ is given in Equation (7.1.17) and

$$\hat{V}(\hat{Y}^{**U}_{ij..}) = \frac{B_{ij}(B_{ij}-b_{ij})}{b_{ij}} s^2_{(\hat{Y}_{ijk.}|k,b_{ij})} + \frac{B_{ij}}{b_{ij}} \sum_{k=1}^{b_{ij}} \hat{V}(\hat{Y}_{ijk.})$$ where $\hat{Y}_{ijk.}$ is given in Equation (7.1.1) and

$$\hat{V}(\hat{Y}_{ijk.}) = \frac{C_{ijk}(C_{ijk}-c_{ijk})}{c_{ijk}} s^2_{(y_{ijk\ell}|\ell,c_{ijk})}$$

(7.1.31) $\hat{\mu} = \hat{\tau}/C...$ where $\hat{\tau}$ is defined by Equation (7.1.29)

(7.1.32) $\hat{V}(\hat{\mu}) = \hat{V}(\hat{\tau})/(C...)^2$ where $\hat{V}(\hat{\tau})$ is defined by Equation (7.1.30).

(7.1.33) $\hat{\tau} = \hat{Y}^{UUU}_{....} = \frac{N}{n} \sum_{i=1}^{n} \hat{Y}^{*UU}_{i...}$ where $\hat{Y}^{*UU}_{i...}$ is given in Equation (7.1.19).

(7.1.34) $\hat{V}(\hat{\tau}) = \frac{N(N-n)}{n} s^2_{(\hat{Y}^{*UU}_{i...}|i,n)} + \frac{N}{n} \sum_{i=1}^{n} \hat{V}(\hat{Y}^{*UU}_{i...})$ where $\hat{Y}^{*UU}_{i...}$ is

given in Equation (7.1.19) and

$$\hat{V}(\hat{Y}^{*UU}_{i...}) = \frac{M(M-m)}{m} s^2_{(\hat{Y}^{**U}_{ij..}|j,m)} + \frac{M}{m} \sum_{j=1}^{m} \hat{V}(\hat{Y}^{**U}_{ij..})$$ where $\hat{Y}^{**U}_{ij..}$ is given

in Equation (7.1.19) and $\hat{V}(\hat{Y}^{**U}_{ij..}) = \frac{B(B-b)}{b} s^2_{(\hat{Y}_{ijk.}|k,b)} + \frac{B}{b} \sum_{k=1}^{b} \hat{V}(\hat{Y}_{ijk.})$

where $\hat{Y}_{ijk.}$ is given in Equation (7.1.3) and $\hat{V}(\hat{Y}_{ijk.}) = \frac{C(C-c)}{c} s^2_{(y_{ijk\ell}|\ell,c}$

(7.1.35) $\hat{\mu} = \hat{\tau}/(NMBC)$ where $\hat{\tau}$ is defined by Equation (7.1.33).

(7.1.36) $\hat{V}(\hat{\mu}) = \hat{V}(\hat{\tau})/(NMBC)^2$ where $\hat{V}(\hat{\tau})$ is defined by Equation (7.1.34).

Bounds on $\hat{V}(\hat{\tau})$ are given by

$$(7.1.37)\quad \hat{V}_L(\hat{\tau}) = T + \frac{N}{n}\sum_{i=1}^{n}\frac{M_i}{m_i}\sum_{j=1}^{m_i}\frac{B_{ij}(B_{ij}-b_{ij})}{b_{ij}}\,s^2(\hat{Y}_{ijk.}\,|\,k,b_{ij})$$

and

$$(7.1.38)\ \text{L1,L5,L6}\quad \hat{V}_U(\hat{\tau}) = T + \frac{N}{n}\sum_{i=1}^{n}\frac{M_i}{m_i}\sum_{j=1}^{m_i}\frac{B_{ij}^2}{b_{ij}}\,s^2(\hat{Y}_{ijk.}\,|\,k,b_{ij})$$ where $\hat{Y}_{ijk.}$ is given in Equation (7.1.1) and

$$T = \frac{N(N-n)}{n}\,s^2(\hat{Y}^{*UU}_{i...}\,|\,i,n) + \frac{N}{n}\sum_{i=1}^{n}\hat{V}_b(\hat{Y}^{*UU}_{i...})$$ where $\hat{Y}^{*UU}_{i...}$ is given in Equation (7.1.17) and

$$\hat{V}_b(\hat{Y}^{*UU}_{i...}) = \frac{M_i(M_i-m_i)}{m_i}\,s^2(\hat{Y}^{**U}_{ij..}\,|\,j,m_i)$$

Bounds on $\hat{V}(\hat{\mu})$ are given by

(7.1.39) $\hat{V}_L(\hat{\mu}) = \hat{V}_L(\hat{\tau})/(C...)^2$ where $\hat{V}_L(\hat{\tau})$ is given by Equation (7.1.37) and

(7.1.40) $\hat{V}_U(\hat{\mu}) = \hat{V}_U(\hat{\tau})/(C...)^2$ where $\hat{V}_U(\hat{\tau})$ is given by Equation (7.1.38).

Bounds on $\hat{V}(\hat{\tau})$ are given by

$$(7.1.41)\quad \hat{V}_L(\hat{\tau}) = T + \frac{N}{n}\sum_{i=1}^{n}\frac{M}{m}\sum_{j=1}^{m}\frac{B(B-b)}{b}\,s^2(\hat{Y}_{ijk.}\,|\,k,b)$$

and

$$(7.1.42)\quad \hat{V}_U(\hat{\tau}) = T + \frac{N}{n}\sum_{i=1}^{n}\frac{M}{m}\sum_{j=1}^{m}\frac{B^2}{b}\,s^2(\hat{Y}_{ijk.}\,|\,k,b)$$ where $\hat{Y}_{ijk.}$ is given in Equation (7.1.3) and

$$T = \frac{N(N-n)}{n} s^2_{(\hat{Y}^{*UU}_{i...}|i,n)} + \frac{N}{n} \sum_{i=1}^{n} \hat{V}_b(\hat{Y}^{*UU}_{i...}) \text{ where } \hat{Y}^{*UU}_{i...} \text{ is given in}$$

Equation (7.1.19) and

$$\hat{V}_b(\hat{Y}^{*UU}_{i...}) = \frac{M(M-m)}{m} s^2_{(\hat{Y}^{**U}_{ij..}|j,m)}$$

Bounds on $\hat{V}(\hat{\mu})$ are given by

(7.1.43) $\hat{V}_L(\hat{\mu}) = \hat{V}_L(\hat{\tau})/(NMBC)^2$ where $\hat{V}_L(\hat{\tau})$ is given by Equation (7.1.41) and

(7.1.44) $\hat{V}_U(\hat{\mu}) = \hat{V}_U(\hat{\tau})/(NMBC)^2$ where $\hat{V}_U(\hat{\tau})$ is given in Equation (7.1.42).

Bounds on $\hat{V}(\hat{\tau})$ are given by

(7.1.45) $\hat{V}_L(\hat{\tau}) = \frac{N(N-n)}{n} s^2_{(\hat{Y}^{*UU}_{i...}|i,n)} + \frac{N}{n} \sum_{i=1}^{n} \hat{V}_L(\hat{Y}^{*UU}_{i...})$ where $\hat{Y}^{*UU}_{i...}$ is given

in Equation (7.1.17) and

$$\hat{V}_L(\hat{Y}^{*UU}_{i...}) = \frac{M_i(M_i-m_i)}{m_i} s^2_{(\hat{Y}^{**U}_{ij..}|j,m_i)} \quad \text{and}$$

(7.1.46) L1,L5,L6 $\hat{V}_U(\hat{\tau}) = \frac{N(N-n)}{n} s^2_{(\hat{Y}^{*UU}_{i...}|i,n)} + \frac{N}{n} \sum_{i=1}^{n} \hat{V}_U(\hat{Y}^{*UU}_{i...})$ where $\hat{Y}^{*UU}_{i...}$ is given

in Equation (7.1.17) and

$$\hat{V}_U(\hat{Y}^{*UU}_{i...}) = \frac{M_i^2}{m_i} s^2_{(\hat{Y}^{**U}_{ij..}|j,m_i)}$$

Bounds on $\hat{V}(\hat{\mu})$ are given by

(7.1.47) $\hat{V}_L(\hat{\mu}) = \hat{V}_L(\hat{\tau})/(C...)^2$ where $\hat{V}_L(\hat{\tau})$ is given by Equation (7.1.45) and

(7.1.48) $\hat{V}_U(\hat{\mu}) = \hat{V}_U(\hat{\tau})/(C...)^2$ where $\hat{V}_U(\hat{\tau})$ is given by Equation (7.1.46).

Bounds on $\hat{V}(\hat{\tau})$ are given by

(7.1.49) $\hat{V}_L(\hat{\tau}) = \frac{N(N-n)}{n} s^2(\hat{Y}^{*UU}_{i...} | i,n) + \frac{N}{n} \sum_{i=1}^{n} V_L(\hat{Y}^{*UU}_{i...})$ where $\hat{Y}^{*UU}_{i...}$ is given in Equation (7.1.19) and

$$\hat{V}_L(\hat{Y}^{*UU}_{i...}) = \frac{M(M-m)}{m} s^2(\hat{Y}^{**U}_{ij..} | j,m) \text{ and}$$

(7.1.50) $\hat{V}_U(\hat{\tau}) = \frac{N(N-n)}{n} s^2(\hat{Y}^{*UU}_{i...} | i,n) + \frac{N}{n} \sum_{i=1}^{n} \hat{V}_U(\hat{Y}^{*UU}_{i...})$ where $\hat{Y}^{*UU}_{i...}$ is given in Equation (7.1.19) and

$$\hat{V}_U(\hat{Y}^{*UU}_{i...}) = \frac{M^2}{m} s^2(\hat{Y}^{**U}_{ij..} | j,m)$$

Bounds on $\hat{V}(\hat{\mu})$ are given by

(7.1.51) $\hat{V}_L(\hat{\mu}) = \hat{V}_L(\hat{\tau})/(NMBC)^2$ where $\hat{V}_L(\hat{\tau})$ is defined by Equation (7.1.49) and

(7.1.52) $\hat{V}_U(\hat{\mu}) = \hat{V}_U(\hat{\tau})/(NMBC)^2$ where $\hat{V}_U(\hat{\mu})$ is defined by Equation (7.1.50).

Bounds on $\hat{V}(\hat{\tau})$ are given by

(7.1.53) $\hat{V}_L(\hat{\tau}) = \frac{N(N-n)}{n} s^2(\hat{Y}^{*UU}_{i...} | i,n)$ where $\hat{Y}^{*UU}_{i...}$ is given in Equation (7.1.17) and

(7.1.54) $\hat{V}_U(\hat{\tau}) = \frac{N^2}{n} s^2(\hat{Y}^{*UU}_{i...} | i,n)$ where $\hat{Y}^{*UU}_{i...}$ is given in Equation (7.1.17)
L1,L6

Bounds on $\hat{V}(\hat{\mu})$ are given by

(7.1.55) $\hat{V}_L(\hat{\mu}) = \hat{V}_L(\hat{\tau})/(C...)^2$ where $\hat{V}_L(\hat{\tau})$ is defined by Equation (7.1.53) and
L1,L6

(7.1.56) $\hat{V}_U(\hat{\mu}) = \hat{V}_U(\hat{\tau})/(C...)^2$ where $\hat{V}_U(\hat{\tau})$ is defined by Equation (7.1.54).
L7

Bounds on $\hat{V}(\hat{\tau})$ are given by

(7.1.57) $\hat{V}_L(\hat{\tau}) = \frac{N(N-n)}{n} s^2(\hat{Y}^{*UU}_{i...} | i,n)$ where $\hat{Y}^{*UU}_{i...}$ is given in Equation (7.1.19) and

(7.1.58) $\hat{V}_U(\hat{\tau}) = \frac{N^2}{n} s^2(\hat{Y}^{*UU}_{i...} | i,n)$ where $\hat{Y}^{*UU}_{i...}$ is given in Equation (7.1.19)

Bounds on $\hat{V}(\hat{\mu})$ are given by

(7.1.59) $\hat{V}_L(\hat{\mu}) = \hat{V}_L(\hat{\tau})/(NMBC)^2$ where $\hat{V}_L(\hat{\tau})$ is defined by Equation (7.1.57)

and

(7.1.60) $\hat{V}_U(\hat{\mu}) = \hat{V}_U(\hat{\tau})/(NMBC)^2$ where $\hat{V}_U(\hat{\tau})$ is defined by Equation (7.1.58).

Chapter 8. SAMPLE SIZE CALCULATIONS FOR MULTI-STAGE DESIGNS.

8.1 Single-Stage Sampling. (Sample Size Formulas) The sample size formulas for cluster sampling are the same as the sample size formulas for simple random sampling. Notice, however, that the mean in SRS corresponds to the mean per cluster in single stage sampling.

Example: Suppose we want the variance of the unbiased estimator of mean per item to be equal to V^*. Then the variance of the unbiased estimator of the mean per cluster equals $V^*(M_o/N)^2 = V$ (say).

Let $S^2 = (N-1)^{-1} \sum_{i=1}^{N} (Y_i - \bar{Y})^2$

where $\bar{Y} = \sum_{i=1}^{N} Y_i/N$ If we ignore the fpc, the sample size n is given by

$n = S^2/V$.

8.2 Two-Stage Sampling. The clusters in two-stage sampling may be of equal size (M) or of unequal size (M_i). Two cases will be treated in this section. The first case is for a two-stage design with clusters of equal size (M) and samples of equal size (m) for sampled clusters. It should be noted in this case that if the only restriction is that nm (total number of sampled second-stage units) is fixed, then n should be chosen as large as possible. The usual constraint, however, involves total cost.

Let the cost function be

(8.2.1) $C = C_1 n + C_2 nm$

where C = total cost,

C_1 = fixed cost per first stage unit (does not vary with m), and

C_2 = cost per second-stage unit (includes measurement and associated cost).

Let $S_u^2 = S_1^2 - \frac{S_2^2}{M}$

where $S_1^2 = \frac{1}{N-1} \sum_{i=1}^{N} (\bar{Y}_{i.} - \bar{Y}_{..})^2$, and

$$S_2^2 = \frac{1}{N(M-1)} \sum_{i=1}^{N} \sum_{j=1}^{M} (Y_{ij} - \bar{Y}_{i.})^2$$

Then, if $S_u^2 \leq 0$, choose m as large as possible. If $S_u^2 > 0$, the optimum value of m, which minimizes cost for a fixed variance (or which minimizes variance for a fixed cost) is given by

$$(8.2.2) \quad m_{opt} = \frac{S_2}{S_u} \sqrt{\frac{C_1}{C_2}}$$

If $m_{opt} > M$, a single-stage design should be used. After obtaining m_{opt}, it is substituted into a variance equation (or cost equation) to obtain n. The variance equation is

$$(8.2.3) \quad V(\hat{\bar{Y}}_{..}) = \frac{N-n}{N} \frac{S_1^2}{n} + \frac{M-m}{M} \frac{S_2^2}{mn}$$

We do not know the values of S_1^2 and S_2^2. A pilot survey can be used to get unbiased estimates. That is,

$$\hat{S}_1^2 = s_1^2 - (\frac{1}{m} - \frac{1}{M}) s_2^2$$

$$\hat{S}_2^2 = s_2^2$$

where $s_1^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\bar{y}_{i.} - \bar{y}_{..})^2$,

$$s_2^2 = \frac{1}{n(m-1)} \sum_{i=1}^{n} \sum_{j=1}^{m} (y_{ij} - \bar{y}_{i.})^2,$$

and n, m, M are sizes used in the pilot survey.

The second case treated in this section is the more general case where clusters are of unequal size (M_i) and samples are of unequal size (m_i). The cost function is similar to the first case but of the form

$$(8.2.4) \quad C = C_1 n + C_2 \sum_{i=1}^{n} m_i$$

$$(8.2.5) \quad E[C] = C_1 n + C_2 \bar{m} n$$

where $\bar{m} = \frac{1}{N} \sum_{i=1}^{N} m_i$

Let $S_A^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(\frac{M_i}{\bar{M}} \bar{Y}_{i.} - \bar{Y}_{..} \right)^2$, and

$$S_w^2 = S_A^2 - \frac{1}{N\bar{M}^2} \sum_{i=1}^{N} M_i S_{2i}^2$$

where $S_{2i}^2 = \frac{1}{M_i - 1} \sum_{j=1}^{M_i} (Y_{ij} - \bar{Y}_{i.})^2$, and $\bar{Y}.. = \frac{1}{N\bar{M}} \sum_{i=1}^{N} M_i \bar{Y}_i$.

We assume that $S_w^2 > 0$. Then, the optimum value of m_i (which minimizes cost for a fixed variance or minimizes variance for a fixed cost--and the variance is that of the unbiased estimator) is given by

$$(8.2.6) \qquad m_i = \sqrt{\frac{C_1}{C_2}} \cdot \frac{M_i}{\bar{M}} \cdot \frac{S_{2i}}{S_w} \text{ for } i = 1, 2, \ldots, N$$

If (8.2.6) is used compute $\bar{m} = \frac{1}{N} \sum_{i=1}^{N} m_i$; substitute into (8.2.5) and solve for n. Again, we must know several things about the population in order to use the formula. When using the unbiased estimator of the mean $\hat{\mu}$, the following approach is more practical.

Let $m_i = f_2 M_i$ for $i = 1, 2, \ldots, N$. This implies $\bar{m} = f_2 \bar{M}$

$$S_2^2 = \frac{1}{M_0} \sum_{i=1}^{N} M_i S_{2i}^2 \text{ where } M. = \sum_{i=1}^{N} M_i$$

Then $\bar{m}_{opt}$ is given by

$$(8.2.7) \qquad \bar{m}_{opt} = \sqrt{\frac{C_1 S_2^2}{C_2 (S_w^2 - S_2^2/\bar{M})}}$$

After obtaining $\bar{m}_{opt}$, it is substituted into the (expected) cost equation (8.2.5) or the variance equation given below to solve for n.

$$(8.2.8) \quad V(\hat{\mu}) = \frac{1-\frac{n}{N}}{n\bar{M}^2} \sum_{i=1}^{N} \frac{(Y_{i.} - \bar{Y})^2}{N-1}$$

$$+ \frac{1-f_2}{n\bar{m}} S_2^2$$

where $f_2 = \bar{m}/\bar{M}$

and $\bar{Y} = \sum_{i=1}^{N} Y_{i.}/N$

8.3 Allocation of Sample to Three Stages for sizes m, M, b, B.

$$V(\hat{\bar{Y}}_{...}) = (\frac{1}{n} - \frac{1}{N}) S_1^2 + \frac{1}{n} (\frac{1}{m} - \frac{1}{M}) S_2^2$$

$$+ \frac{1}{nm} (\frac{1}{b} - \frac{1}{B}) S_3^2$$

$$= - \frac{1}{N} S_1^2 + \frac{1}{n} S_u^2 + \frac{1}{nm} S_v^2 + \frac{1}{nmb} S_3^2$$

where $S_1^2 = \sum_{i=1}^{N} (\bar{Y}_{i..} - \bar{Y}_{...})^2/(N-1)$

$$S_2^2 = \sum_{i=1}^{N} \sum_{j=1}^{M} (\bar{Y}_{ij.} - \bar{Y}_{i..})^2/(N(M-1))$$

$$S_3^2 = \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k=1}^{B} (Y_{ijk} - \bar{Y}_{ij.})^2/(NM(B-1))$$

Let $S_u^2 = S_1^2 - S_2^2/M$ and $S_v^2 = S_2^2 - S_3^2/B$.

We assume a cost function of the form $C = C_1 n + C_2 nm + C_3 nmb$

Case I $S_u^2 > 0$, $S_v^2 > 0$

Optimum sample size which minimizes cost for fixed variance or minimizes variance for fixed cost is given by

$$m_{opt} = \left[\frac{C_1 \, S_v^2}{C_2 \, S_u^2} \right]^{1/2}$$

and

$$b_{opt} = \left[\frac{C_2 \, S_3^2}{C_3 \, S_v^2} \right]^{1/2}$$

Case II $S_u^2 \leq 0$, but $S_v^2 > 0$

Choose m as large as possible (which can also be equal to M).

If $C_3 \, S_v^2 + C_3 \, m_{opt} \, S_u^2 \leq 0$, choose b as large as possible. Otherwise, choose b to be the integer closest to

$$\left[\frac{(\frac{1}{m_{opt}} C_1 + C_2) \, S_3^2}{C_3 \, S_v^2 + m_{opt} C_3 \, S_u^2} \right]^{1/2}$$

Case III $S_u^2 \leq 0$, $S_v^2 \leq 0$

Choose m and b to be their respective maximum attainable values.

Case IV $S_u^2 > 0$ but $S_v^2 \leq 0$

For this case, solve for K where given by

$$K = \left[\frac{C_1 \, S_3^2}{C_3 \, S_u^2} \right]^{1/2}$$

Then choose the minimum possible m and the maximum possible b such that $bm = K$.

Note. In practice, we do not know S_1^2, S_2^2 and S_3^2. If we do a pilot survey with sample sizes n, m, b, we can unbiasedly estimate S_1^2, S_2^2 and S_3^2 using the following formulas.

$$\hat{S}_1^2 = s_1^2 - \frac{(1-m/M) \, S_2^2}{m} - \frac{(1-b/B) \, s_3^2}{mb}$$

$$\hat{S}_2^2 = s_2^2 - \frac{(1-b/B)\ s_3^2}{b}$$

$$\hat{S}_3^2 = s_3^2$$

where s_1^2, s_2^2 and s_3^2 are sample values corresponding to S_1^2, S_2^2 and S_3^2 respectively.

8.4 Allocation of Sample to Four-Stages for Sizes m, M, b, B, c, C.

The variance of the unbiased estimator of the mean is given by

(8.4.1) $$V(\hat{\bar{Y}}_{....}) = (\frac{1}{n} - \frac{1}{N})\ S_1^2 + \frac{1}{n}\ (\frac{1}{m} - \frac{1}{M})\ S_2^2$$

$$+ \frac{1}{nm}\ (\frac{1}{b} - \frac{1}{B})\ S_3^2 + \frac{1}{nmb}\ (\frac{1}{c} - \frac{1}{C})\ S_4^2$$

where $$S_1^2 = \sum_{i=1}^{N} (\bar{Y}_{i...} - \bar{Y}_{....})^2/(N-1)$$

$$S_2^2 = \sum_{i=1}^{N} \sum_{j=1}^{M} (\bar{Y}_{ij..} - \bar{Y}_{i...})^2/(N(M-1))$$

$$S_3^2 = \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k=1}^{B} (\bar{Y}_{ijk.} - \bar{Y}_{ij..})^2/(NM(B-1))$$

$$S_4^2 = \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k=1}^{B} \sum_{\ell=1}^{C} (Y_{ijk\ell} - \bar{Y}_{ijk.})^2/(NMB(C-1))$$

Let $$S_u^2 = S_1^2 - S_2^2/M$$

$$S_v^2 = S_2^2 - \frac{S_3^2}{B}$$

$$S_w^2 = S_3^2 - \frac{S_4^2}{C}$$

We assume that $S_u^2 > 0$, $S_v^2 > 0$ and $S_w^2 > 0$

Let the cost function be of the form

(8.4.2) $\quad \text{Cost} = C_1 n + C_2 nm + C_3 nmb + C_4 nmbc$

The optimum values which minimize cost for fixed variance (or minimize variance for fixed cost) are given by

$$(8.4.3)\quad m_{opt} = \frac{S_v}{S_u}\sqrt{\frac{C_1}{C_2}}$$

$$(8.4.4)\quad b_{opt} = \frac{S_w}{S_v}\sqrt{\frac{C_2}{C_3}}$$

$$(8.4.5)\quad c_{opt} = \frac{S_4}{S_w}\sqrt{\frac{C_3}{C_4}}$$

To get the value of n_{opt} substitute into the cost equation (8.4.2) or variance equation (8.4.1). To utilize the sample size formulas estimates of the S_i^2 would be used. These estimates could be obtained from a pilot study of sizes n, m, b, c.

Unbiased estimates of S_i^2 (i = 1, 2, 3, 4) are given by

$$\hat{S}_1^2 = s_1^2 - \frac{(1-m/M)\ \hat{S}_2^2}{m} - \frac{(1-b/B)}{mb}\hat{S}_3^2 - \frac{(1-c/C)}{mbc}s_4^2$$

$$\hat{S}_2^2 = s_2^2 - \frac{(1-b/B)}{b}\hat{S}_3^2 - \frac{(1-c/C)}{bc}s_4^2$$

$$\hat{S}_3^2 = s_3^2 - \frac{(1-c/C)}{c}s_4^2$$

$$\hat{S}_4^2 = s_4^2$$

where s_1^2, s_2^2, s_3^2 and s_4^2 are sample values corresponding to S_1^2 S_2^2, S_3^2 and S_4^2 respectively.

Chapter 9. ONE - AND TWO-PHASE SAMPLING DESIGNS

9.1 Introduction. Multi-phase designs allow use of information in auxiliary variables from various phases (levels) in estimation of population parameters. Estimators involving auxiliary variables can be constructed in many ways; we shall restrict our discussion of the use of auxiliary variables to multi-phase sampling for stratification. At each phase of sampling (except the last) an auxiliary variable will be measured to classify the unit into a stratum. Thus, the types of designs we shall present in PART III are very similar to stratified sampling. The only difference is that we shall be using estimated stratum sizes rather than known stratum sizes.

Regardless of the number of levels (phases) of information used in the design, the size of the "sample units" remains the same. That is to say, a first-phase unit is the same size as a second-phase unit, etc. This is in contrast to the multi-stage designs discussed in PART II where each succeeding stage involved a partitioning of units into smaller units.

For each multi-phase design discussed, equations will be given for $\hat{\mu}$ and $\hat{V}(\hat{\mu})$. To get the corresponding estimates for population totals, the following equations are used.

(9.1.1) $\hat{\tau} = N\hat{\mu}$, and

(9.1.2) $\hat{V}(\hat{\tau}) = N^2\hat{V}(\hat{\mu})$

where N is the population size.

It should be mentioned here that sampling fractions for strata in the various phases (as later designated by u_i, v_{ij}, z_{ijk}) are fixed in advance. Regardless of how the values of the sampling fractions are chosen, (subject to the restrictions given later) the equations for $\hat{\mu}$, $\hat{\tau}$, $\hat{V}(\hat{\mu})$ and $\hat{V}(\hat{\tau})$ are valid.

PART III

MULTI-PHASE SAMPLING DESIGNS

9.2 Single-Phase Sampling. Single- or one-phase sampling is a special case of multi-phase sampling. It is actually simple random sampling. Equations and examples are given in Chapter 2 with additional examples given in Chapter 4.

9.3 Two-Phase Sampling. Two-phase sampling, also called double sampling, is used frequently in natural resource inventories (Bickford, et al; 1961). As an example, when sampling a large area such as a state, a sample of first-phase units can be selected at random and located on aerial photographs. These units can be stratified according to estimated land use classes, estimated vegetative cover classes, or any other classification method. A subsample of the first-phase units can then be selected and visited on the ground to obtain population values such as timber volume, range productivity, etc. This approach has proven highly satisfactory in many instances. In most applications, the stratification has led to more precise estimates than available with simple random sampling due to a high ratio of cost of second-phase units to cost of first-phase units.

The general procedure involves selecting a simple random sample of size n of the N population units in the first phase. The n units are then stratified into I strata. We then designate the number of first-phase sample units in stratum i as n_i, and

$$n = \sum_{i=1}^{I} n_i$$

The unknown total number of units in stratum i is designated as N_i, and

$$N = \sum_{i=1}^{I} N_i$$

A number (sampling fraction), u_i, is then specified for stratum i, and the resulting number of units in the second-phase sample (subsample) in stratum i is designated as m_i and is calculated by

(9.3.1) $$m_i = u_i n_i;\ i = 1,2,\ldots,I$$

To derive the formulas given in the equations, the following assumptions are made:

(i) $0 < u_i < 1$,

(ii) m_i's are integers,

(iii) $P[m_i \geq 2] = 1$, and

(iv) the m_i units in stratum i are selected by simple random sampling.

In a practical situation, m_i in (9.3.1) is rounded to the closest integer greater than or equal to two. The characteristic of interest (population value) is then measured and designated as $y_{ij};\ j = 1,2,\ldots,m_i;\ i = 1,2,..,I$

The unbiased point estimate of the population mean is given by

(9.3.2) $$\hat{\mu} = \sum_{i=1}^{I} w_i \bar{y}_i$$

where $$w_i = \frac{n_i}{n}, \text{ and}$$

$$\bar{y}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij}$$

The proofs for (9.3.2) and subsequent equations in this section can be found in Cochran (1976).

The variance of $\hat{\mu}$ is given by

(9.3.3) $$V(\hat{\mu}) = S^2\left(\frac{1}{n} - \frac{1}{N}\right) + \sum_{i=1}^{I} \left[\frac{W_i S_i^2}{n}\left(\frac{1}{u_i} - 1\right)\right]$$

where S^2 is the population variance,

S_i^2 is the variance of the N_i units in stratum i, and

$W_i = \frac{N_i}{N}$

As usual there are the unknown parameters in (9.3.3), so we must resort to an estimate. The unbiased estimate of the variance of the estimated population mean is given by

$$(9.3.4)\quad \hat{V}(\hat{\mu}) = \frac{N-1}{N} \sum_{i=1}^{I} \left[\left(\frac{n_i - 1}{n-1} - \frac{m_i - 1}{N-1} \right) \frac{w_i s_i^2}{m_i} \right] + \frac{N-n}{N(n-1)} \sum_{i=1}^{I} w_i (\bar{y}_i - \hat{\mu})^2$$

$$(9.3.5)\quad \text{where} \quad s_i^2 = \frac{1}{m_i - 1} \sum_{j=1}^{m_i} (y_{ij} - \bar{y}_i)^2$$

Sample size calculations are discussed in Chapter 12.

Multiphase examplesgiven here and those in succeeding chapters were developed with two basic assumptions. First the cost assumption: $100.00 to measure one 1.1 - acre ground plot; $5.00 to classify one 1.1 - acre low altitude photo plot; $2.00 to classify one 1.1 - acre high altitude photo plot; $1.00 to classify one 1.1 - acre landsat pixel. The second assumption dealt with the ratio of measured (classified) plots at each phase of imagery. For every 100 pixels classified, 10 high altitude photo plots were classified. For every 10 high altitude photo plots classified, five low altitude photo plots were classified, and for every five low altitude photo plots classified, one ground plot was measured. The cost and ratio assumptions lead to approximately equal costs for each example. The ratio assumption also determined the sampling fractions used in the examples.

Example 9.3.1 - Given a universe of N = 1,016,064 1.1 - acre landsat pixels, 1008 pixels on aside, a simple random sample (without replacement) of n = 50,400 pixels was chosen at phase 1. These pixels were grouped into seven strata: conifer, hardwood, sage, other brush, pasture/meadow, grass/barren, and water. The following were the phase 1 sample sizes: $n_1 = 20,296$, $n_2 = 10,175$, $n_3 = 6,940$, $n_4 = 3,132$, $n_5 = 3,082$, $n_6 = 6,116$, and $n_7 = 659$. A sampling fraction of u = 0.01 was selected to collect a simple random phase 2 sample (without replacement) of 1.1 acre ground plots from each of the above strata. The following were the phase 2 sample sizes: $m_1 = 203$, $m_2 = 102$, $m_3 = 69$, $m_4 = 31$, $m_5 = 31$, $m_6 = 61$, and $m_7 = 7$. The data collected at phase 2 were used to estimate average timber volume per 1.1 - acre ($\hat{\mu}$ timber volume), average herbage per 1.1 - acre ($\hat{\mu}$ herbage), and the proportion of forest area ($\hat{p}$ forest). The results were:

	$\hat{\theta}$	$\sqrt{\hat{V}(\hat{\theta})}$	$100\sqrt{\hat{V}(\hat{\theta})}/\hat{\theta}$
timber volume, $\hat{\theta} = \hat{\mu}$ timber volume	991.1	49.8	5.0%
herbage, $\hat{\theta} = \hat{\mu}$ herbage	1176.0	60.2	5.1%
forest area, $\hat{\theta}$ $\hat{p}$ forest	0.5175	0.0141	2.7%

Example 9.3.2. - Given a universe of N = 1,016,064 1.1 - acre high altitude photo plots, 1008 plots on a side, a simple random sample (without replacement) of n = 8400 photo plots was chosen at phase 1. These plots were grouped into seven strata : conifer, hardwood, sage, other brush, pasture/meadow, grass/barren, and water. The following were the phase 1 sample sizes: $n_1 = 2,830$, $n_2 = 1,558$, $n_3 = 1,401$, $n_4 = 650$, $n_5 = 920$, $n_6 = 856$, and $n_7 = 185$. A sampling fraction of u = 0.1 was selected to collect a simple random phase 2 sample (without replacement) of 1.1 - acre ground plots from each of the above strata. The following were the phase 2 sample sizes: $m_1 = 283$, $m_2 = 156$, $m_3 = 140$, $m_4 = 65$, $m_5 = 92$, $m_6 = 86$, and $m_7 = 19$. The data collected at phase 2 were used to estimate average timber volume per 1.1 - acre ($\hat{\mu}$ timber volume), average herbage per 1.1 - acre ($\hat{\mu}$ herbage), and the proportion of forest area ($\hat{p}$ forest). The results were:

	$\hat{\theta}$	$\sqrt{\hat{V}(\hat{\theta})}$	$100\sqrt{\hat{V}(\hat{\theta})}/\hat{\theta}$
timber volume, $\hat{\theta} = \hat{\mu}$ timber volume	971.5	36.6	3.6%
herbage $\hat{\theta}$ $\hat{\mu}$ herbage	1247.7	43.1	3.5%
forest area, $\hat{\theta}$ $\hat{p}$ forest	0.4915	0.0100	2.0%

Example 9.3.3. Given a universe of N = 1,016,064 1.1 - acre low altitude photo plots, 1008 plots on a side, a simple random sample (without replacement) of n = 4030 photo plots was chosen at phase 1. These plots were grouped into seven strata: conifer, hardwood, sage, other brush, pasture/meadow, grass/barren, and water. The following were the phase 1 sample sizes: $n_1 = 1{,}059$, $n_2 = 1{,}062$, $n_3 = 544$, $n_4 = 307$, $n_5 = 516$, $n_6 = 431$, $n_7 = 111$. A sampling fraction of u = 0.2 was selected to collect a simple random phase 2 sample (without replacement) of 1.1 - acre ground plots from each of the above strata. The following were the phase 2 sample sizes: $m_1 = 212$, $m_2 = 212$, $m_3 = 109$, $m_4 = 61$, $m_5 = 103$, $m_6 = 86$, $m_7 = 22$. The data collected at phase 2 were used to estimate average timber volume per 1.1 - acre ($\hat{\mu}$ timber volume), average herbage per 1.1 - acre ($\hat{\mu}$ herbage), and the proportion of forest area ($\hat{p}$ forest). The results were:

	$\hat{\theta}$	$\sqrt{\hat{V}(\hat{\theta})}$	$100\sqrt{\hat{V}(\hat{\theta})}/\hat{\theta}$
timber volume, $\hat{\theta} = \hat{\mu}$ timber volume	994.8	37.8	3.8%
herbage, $\hat{\theta} = \hat{\mu}$ herbage	1216.9	42.9	3.5%
forest area, $\hat{\theta} = \hat{p}$ forest	0.5002	0.0100	2.0%

Chapter 10. THREE-PHASE SAMPLING DESIGNS.

10.1 Estimation Procedures. Unbiased estimators for variances of estimated population means have not been available for ℓ-phase sampling ($\ell \geq 3$). One of the most notable accomplishments of this project is development of such estimators. The derivation of estimators used in Chapters 10 and 11 is covered in Appendix I.

We shall use the same notation and assumptions in this Chapter as used in Chapter 9 with some changes and additions:

J_i = the number of phase-two strata in phase-one stratum i, $i = 1,2,\ldots,I$;

m_{ij} = the number of units in m_i included in phase-two stratum j; $j = 1,2,\ldots,J_i$; $i = 1,2,\ldots,I$; where $m_i = \sum_{j=1}^{J_i} m_{ij}$

A sampling fraction v_{ij} is specified for stratum ij. The resulting number of units in the third-phase sample in stratum ij is designated as b_{ij} and is calculated by

$$b_{ij} = v_{ij} m_{ij} \tag{10.1.1}$$

where the b_{ij}'s are integers,

$$0 < v_{ij} \leq 1,$$

$$p[b_{ij} \geq 2] = 1, \text{ and}$$

the b_{ij} units in stratum ij are selected by simple random sampling.

The characteristic of interest, y_{ijk}, is then measured; $k = 1,2,\ldots,b_{ij}$; $j = 1,2,\ldots,m_i$; $i = 1,2,\ldots,I$

The unbiased point estimate of the population mean is given by

$$\hat{\mu} = \sum_{i=1}^{I} w_i \sum_{j=1}^{J_i} w_{ij} \bar{y}_{ij} = \sum_{i=1}^{I} w_i \bar{y}_i \tag{10.1.2}$$

where $w_{ij} = \frac{m_{ij}}{m_i}$,

$$\bar{y}_{ij} = \frac{1}{b_{ij}} \sum_{k=1}^{b_{ij}} y_{ijk}, \text{ and}$$

$$\bar{y}_i = \sum_{j=1}^{J_i} \frac{m_{ij}}{m_i} \bar{y}_{ij}$$

The unbiased estimate of the variance of the estimated population mean is given by

$$\text{(10.1.3)} \qquad \hat{V}(\hat{\mu}) = (1-h)^{-1} \left\{ h \sum_{i=1}^{I} w_i (\bar{y}_i - \hat{\mu})^2 + \frac{h}{n} \sum_{i=1}^{I} (n_i - 1) s_i^{2*} + \sum_{i=1}^{I} [w_i (w_i - h) \hat{V}_i] \right\}$$

where $h = \dfrac{N-n}{n(N-1)}$

$$\text{(10.1.4)} \qquad s_i^{2*} = \frac{\hat{V}_i - \dfrac{1}{m_i} \sum_{j=1}^{J_i} w_{ij} s_{ij}^2 \left(\dfrac{m_{ij}}{b_{ij}} - 1 \right)}{\dfrac{1}{m_i} - \dfrac{1}{n_i}}$$

$$\text{(10.1.5)} \qquad \hat{V}_i = \frac{m_i (n_i - 1)}{(m_i - 1) n} \left\{ \sum_{j=1}^{J_i} w_{ij} s_{ij}^2 \left(\frac{1}{m_i v_{ij}} - \frac{1}{n_i} \right) + \frac{n_i - m_i}{m_i (n_i - 1)} \left[\sum_{j=1}^{J_i} s_{ij}^2 \left(\frac{w_{ij}}{n_i} - \frac{1}{m_i v_{ij}} \right) + \sum_{j=1}^{J_i} w_{ij} (\bar{y}_{ij} - \bar{y}_i)^2 \right] \right\}$$

$$\text{(10.1.6)} \qquad s_{ij}^2 = \frac{\sum_{k=1}^{b_{ij}} (y_{ijk} - \bar{y}_{ij})^2}{b_{ij} - 1}$$

An alternative computational equation for $\hat{V}_i$ as given by Cochran (1976) is

$$\text{(10.1.7)} \qquad \hat{V}_i = \frac{n_i - 1}{n_i} \sum_{j=1}^{J_i} \left(\frac{m_{ij} - 1}{m_i - 1} - \frac{b_{ij} - 1}{n_i - 1} \right) \frac{w_{ij} s_{ij}^2}{b_{ij}} + \frac{n_i - m_i}{n_i (m_i - 1)} \sum_{j=1}^{J_i} w_{ij} (\bar{y}_{ij} - \bar{y}_i)^2$$

Note. s_i^{2*} is an unbiased estimate of the variance of the n_i units (Cochran, 1976).

Note. $E(\bar{y}_i | n_i \text{ units}) = \bar{Y}_{i\,(n_i)} = \text{mean of the } n_i \text{ units.}$

$\hat{V}_i$ is an unbiased estimate of the conditional variance of $\bar{y}_i$ given the n_i units (Cochran, 1976).

Example 10.1.1 - Given a universe of N = 1,016,064 1.1 - acre landsat pixels, 1008 pixels on a side, a simple random sample (without replacement) of n = 45,800 pixels was chosen at phase 1. These pixels were grouped into three strata: conifer/ hardwood, water, and all other land classes, with the following phase 1 sample sizes; n_1 = 27,752, n_2 = 603, and n_3 = 17,445. A sampling fraction of u = 0.1 was selected to collect a simple random phase 2 sample (without replacement) of 1.1 - acre high altitude photoplots from each of the above phase 1 strata. The phase 2 sample sizes were: m_1 = 2,775, m_2 = 60, and m_3 = 1,745. At phase 2, each of the above strata was further stratified. The phase 1 conifer/hardwood stratum was grouped into three phase 2 strata: conifer, hardwood, and all other land classes. The phase 1 water stratum was grouped into one phase 2 stratum incorporating all land classes, and the phase 1 "all other" stratum was grouped into two phase 2 strata: conifer/hardwood, and all other land classes. The resulting phase 2 stratified sample sizes were: m_{11} = 1,553, m_{12} = 758, m_{13} = 464, m_{21} = 60, m_{31} = 101, and m_{32} = 1644. A sampling fraction of v = 0.1 was selected to collect a simple random phase 3 sample (without replacement) of 1.1 - acre ground plots from each of the above phase 2 strata. The following were the phase 3 sample sites: b_{11} = 155, b_{12} = 76, b_{13} = 46, b_{21} = 6, b_{31} = 10, and b_{32} = 164. The data collected at phase 3 were used to estimate average timber volume per 1.1 -acre ($\hat{\mu}$ timber volume), average herbage per 1.1 - acre ($\hat{\mu}$ herbage), and the proportion of forest area ($\hat{p}$ forest). The results were:

	$\hat{\theta}$	$\sqrt{\hat{V}(\hat{\theta})}$	$100\sqrt{\hat{V}(\hat{\theta})}/\hat{\theta}$
timber volume, $\hat{\theta} = \hat{\mu}$ timber volume	1094.9	50.9	4.6%
herbage, $\hat{\theta} = \hat{\mu}$ herbage	1149.5	53.6	4.7%
forest area, $\hat{\theta} = \hat{p}$ forest	0.5113	0.0079	1.5%

Example 10.1.2 - Given a universe of N = 1,016,064 1.1 - acre landsat pixels, 1008 pixels on a side, a simple random sample (without replacement) of n = 44,800 pixels was chosen at phase 1. These pixels were grouped into three strata: conifer/hardwood, water, and all other land classes, with the following phase 1 sample sizes; n_1 = 27,147, n_2 = 590, and n_3 = 17,063. A sampling fraction of u = 0.05 was selected to collect a simple random phase 2 sample (without replacement) of 1.1 - acre low altitude photoplots from each of the above phase 1 strata. The phase 2 sample sizes were: m_1 = 1357, m_2 = 30, and m_3 = 853. At phase 2, each of the above strata was further stratified. The phase 1 conifer/hardwood stratum was grouped into three phase 2 strata: conifer, hardwood, and all other land classes. The phase 1 water stratum was grouped into one phase 2 stratum incorporating all land classes, and the phase 1 "all other" stratum was grouped into two phase 2 strata: conifer/hardwood, and all other land classes. The resulting phase 2 stratified sample sizes were: m_{11} = 574, m_{12} = 549, m_{13} = 234, m_{21} = 30, m_{31} = 50, m_{32} = 803. A sampling fraction of v = 0.2 was selected to collect a simple random phase 3 sample (without replacement) of 1.1 - acre ground plots from each of the above phase 2 strata. The following were the phase 3 sample sizes: b_{11} = 115, b_{12} = 110, b_{13} = 47, b_{21}= 6, b_{31} = 10, and b_{32} = 161. The data collected at phase 3 were used to estimate average timber volume per 1.1 - acre ($\hat{\mu}$ timber volume), average herbage per 1.1 - acre ($\hat{\mu}$ herbage), and the proportion of forest area ($\hat{p}$ forest). The results were:

	$\hat{\theta}$	$\sqrt{\hat{V}(\hat{\theta})}$	$100\sqrt{\hat{V}(\hat{\theta})}/\hat{\theta}$
timber volume, $\hat{\theta} = \hat{\mu}$ timber volume	1073.8	49.3	4.6%
herbage, $\hat{\theta} = \hat{\mu}$ herbage	1171.9	52.6	4.5%
forest area, $\hat{\theta} = \hat{p}$ forest	0.5037	0.0095	1.9%

Example 10.1.3. - Given a universe of N = 1,016,064 1.1 - acre high altitude photoplots, 1008 plots on aside, a simple random sample (without replacement) of n = 6,950 plots was chosen at phase 1. These plots were grouped into three strata: conifer/hardwood, water, and all other land classes, with the following phase 1 sample sizes; n_1 = 3,647, n_2 = 160, n_3 = 3,143. A sampling fraction of u = 0.5 was selected to collect a simple random phase 2 sample (without replacement) of 1.1 - acre low altitude photoplots from each of the above phase 1 strata. The phase 2 sample sizes were: m_1 = 1,824, m_2 = 80, and m_3 = 1572. At phase 2, each of the above strata was further stratified. The phase 1 conifer/hardwood stratum was grouped into three phase 2 strata: conifer, hardwood, and all other land classes. The phase 1 water stratum was grouped into one phase 2 stratum incorporating all land classes, and the phase 1 "all other" stratum was grouped into two phase 2 strata, conifer/hardwood and all other land classes. The resulting phase 2 stratified sample sizes were: m_{11} = 898, m_{12} = 894, m_{13} = 32, m_{21} = 80, m_{31} = 14, and m_{32} = 1558. A sampling fraction of v = 0.2 was selected to collect a simple random phase 3 sample (without replacement) of 1.1 - acre ground plots from each of the above phase 2 strata. The following were the phase 3 sample sizes; b_{11} = 180, b_{12} = 179, b_{13} = 6, b_{21} = 16, b_{31} = 3, and b_{32} = 312. The data collected at phase 3 were used to estimate average timber volume per 1.1 - acre ($\hat{\mu}$ timber volume), average herbage per 1.1 - acre ($\hat{\mu}$ herbage), and the proportion of forest area ($\hat{p}$ forest). The results were:

	$\hat{\theta}$	$\sqrt{\hat{V}(\hat{\theta})}$	$100\sqrt{\hat{V}(\hat{\theta})}/\hat{\theta}$
timber volume, $\hat{\theta}$ = $\hat{\mu}$ timber volume	1077.4	41.2	3.8%
herbage, $\hat{\theta}$ = $\hat{\mu}$ herbage	1264.6	50.8	4.0%
forest area, $\hat{\theta}$ = $\hat{p}$ forest	0.4937	0.0083	1.7%

Chapter 11. FOUR-PHASE SAMPLING DESIGNS.

11.1 Estimation Procedures. We shall use the same notation and assumptions in this Chapter as used in Chapter 10 with some changes and additions:

K_{ij} = the number of phase three strata in phase two stratum j in phase one stratum i, $j = 1,2,\ldots,J_i$; $i = 1,2,\ldots,I$;

b_{ijk} = the number of units in b_{ij} included in phase three stratum k; $k = 1,2,\ldots,K_{ij}$; $j = 1,2,\ldots,J_i$; $i = 1,2,\ldots,I$;

where $b_{ij} = \sum_{k=1}^{K_{ij}} b_{ijk}$

A sampling fraction z_{ijk} is specified for stratum ijk. The resulting number of units in the fourth-phase sample in stratum ijk is designated as c_{ijk} and is calculated by

(11.1.1) $c_{ijk} = z_{ijk} b_{ijk}$

where the c_{ijk}'s are integers,

$0 < z_{ijk} < 1$,

$P[c_{ijk} \geq 2] = 1$, and

the c_{ijk} units in stratum ijk are selected by simple random sampling.

The characteristic of interest, $y_{ijk\ell}$, is then measured; $\ell = 1,2,\ldots,c_{ijk}$; $k = 1,2,\ldots,b_{ij}$; $j = 1,2,\ldots,m_i$; $i = 1,2,\ldots,I$

The unbiased point estimate of the population mean is given by

(11.1.2) $$\hat{\mu} = \sum_{i=1}^{I} w_i \sum_{j=1}^{J_i} w_{ij} \sum_{k=1}^{K_{ij}} w_{ijk} \bar{y}_{ijk} = \sum_{i=1}^{I} w_i \hat{\bar{y}}_i$$

where $w_{ijk} = \dfrac{b_{ijk}}{b_{ij}}$,

$$\bar{y}_{ijk} = \frac{1}{c_{ijk}} \sum_{\ell=1}^{c_{ijk}} y_{ijk\ell},$$

$$\hat{\bar{y}}_i = \sum_{j=1}^{J_i} w_{ij} \hat{\bar{y}}_{ij}, \text{ and}$$

$$\hat{\bar{y}}_{ij} = \sum_{k=1}^{K_{ij}} w_{ijk} \bar{y}_{ijk}$$

The unbiased estimate of the variance of the estimated population mean is given by

$$(11.1.3) \quad \hat{V}(\hat{\mu}) = (1-h)^{-1} \left\{ h \sum_{i=1}^{I} w_i (\hat{\bar{y}}_i - \hat{\mu})^2 + \frac{h}{n} \sum_{i=1}^{I} (n_i - 1) s_i^{2*} + \sum_{i=1}^{I} [w_i (w_i - h) \hat{V}_i] \right\}$$

Note. (11.1.3) appears to be the same as (10.1.3), (this equation is the same for any multi-phase design) but the quantities in the formula depend on the number of phases.

$$(11.1.4) \quad \hat{V}_i = \frac{m_i}{m_i - f_i} \left\{ \frac{f_i}{m_i} \left[\sum_{j=1}^{J_i} w_{ij} (\hat{\bar{y}}_{ij} - \hat{\bar{y}}_i)^2 + \sum_{j=1}^{J_i} (w_{ij} - \frac{1}{m_i}) s_{ij}^{2*} \right] + \sum_{j=1}^{J_i} w_{ij} (w_{ij} - \frac{f_i}{m_i}) \hat{V}_{ij} \right\}$$

where $f_i = \dfrac{n_i - m_i}{n_i - 1}$

$$(11.1.5) \quad \hat{V}_{ij} = \frac{m_{ij} - 1}{m_{ij}} \sum_{k=1}^{K_{ij}} \left[\left(\frac{b_{ijk} - 1}{b_{ij} - 1} - \frac{c_{ijk} - 1}{m_{ij} - 1} \right) \frac{w_{ijk}}{c_{ijk}} s_{ijk}^2 \right] + \frac{m_{ij} - b_{ij}}{m_{ij}(b_{ij} - 1)} \sum_{k=1}^{K_{ij}} w_{ijk} (\hat{\bar{y}}_{ijk} - \hat{\bar{y}}_{ij})^2$$

$$(11.1.6) \quad s_{ijk}^2 = \frac{1}{c_{ijk} - 1} \sum_{\ell=1}^{c_{ijk}} (y_{ijk\ell} - \hat{\bar{y}}_{ijk})^2$$

11.1.7 $$s_{ij}^{2}* = \frac{\hat{V}_{ij} - \frac{1}{b_{ij}} \sum_{k=1}^{K_{ij}} w_{ijk} \left(\frac{b_{ijk}}{c_{ijk}} - 1\right) s_{ijk}^{2}}{\left(\frac{1}{b_{ij}} - \frac{1}{m_{ij}}\right)}$$

11.1.8 $$s_{i}^{2}* = \frac{\hat{V}_{i} - \sum_{j=1}^{J_i} w_{ij}^{2} \hat{V}_{ij}}{\frac{1}{m_i} - \frac{1}{n_i}}$$

Example 4.1.1.- Given a universe of N = 1,016,064 1.1 - acre landsat pixels, 1008 pixels on a side, a simple random sample (without replacement) of n = 41,100 pixels was chosen at phase 1. These pixels were grouped into three strata: conifer/hardwood, water, and all other land classes, with the following phase 1 sample sizes: n_1 = 24,927, n_2 = 548, n_3 = 15,625. A sampling fraction of u = 0.1 was selected to collect a simple random phase 2 sample (without replacement) of 1.1 - acre high altitude photoplots from each of the above phase 1 strata. The phase 2 sample sizes were: m_1 = 2493, m_2 = 55, m_3 = 1563. At phase 2, each of the above strata was further stratified. The phase 1 conifer/hardwood stratum was grouped into two phase 2 strata: conifer/hardwood and "all other" land classes. The phase 1 water stratum was grouped into one phase 2 stratum incorporating all land classes. The phase 1 "all other" stratum also grouped into one phase 2 stratum incorporating all land classes. The resulting phase 2 stratified sample sizes were: m_{11} = 2,071, m_{12} = 422, m_{21} = 55, and m_{31} = 1563. A sampling fraction of v = 0.5 was selected to collect a simple random phase 3 sample (without replacement) of 1.1 - acre low altitude photoplots from each of the above phase 2 strata. The following were the phase 3 sample sizes: b_{11} = 1036, b_{12} = 211, b_{21} = 28, and b_{31} = 782. At phase 3, each of the above phase 2 strata was further stratified. The phase 2 conifer/hardwood stratum was grouped into three phase 3 strata; conifer, hardwood, and all other land classes. The phase 2 "all other"stratum was grouped into one phase 3 stratum incorporating all land classes. The first "all land classes" phase 2 stratum (obtained from the phase 1 water stratum) was grouped into one phase 3 stratum incorporating all land classes, and the second"all land classes" phase 2 stratum (obtained from the phase 1 "all other" stratum) was grouped into two phase 3 strata: conifer/hardwood and "all other" land classes. The resulting phase 3 stratified sample sizes were: b_{111} = 530, b_{112} = 489, b_{113} = 17, b_{121} = 211,

$b_{211} = 28$, $b_{311} = 53$, and $b_{312} = 729$. A sampling fraction of $z = 0.2$ was selected to collect a simple random phase 4 sample (without replacement) of 1.1 - acre ground plots from each of the above phase 3 strata. The following were the phase 4 sample sizes: $c_{111} = 106$, $c_{112} = 98$, $c_{113} = 3$, $c_{121} = 42$, $c_{211} = 6$, $c_{311} = 11$, and $c_{312} = 146$. The data collected at phase 4 were used to estimate average timber volume per 1.1 - acre ($\hat{\mu}$ timber volume), average herbage per 1.1 -acre ($\hat{\mu}$ herbage), and the proportion of forest area ($\hat{p}$ forest). The results were:

	$\hat{\theta}$	$\sqrt{\hat{V}(\hat{\theta})}$	$100\sqrt{\hat{V}(\hat{\theta})}/\hat{\theta}$
timber volume, $\hat{\theta} = \hat{\mu}$ timber volume	1109.9	57.8	5.2%
herbage, $\hat{\theta}$ $\hat{\mu}$ herbage	1237.6	62.6	5.1%
forest area, $\hat{\theta}$ $\hat{p}$ forest	0.4948	0.0097	2.0%

Chapter 12. SAMPLE SIZE CALCULATIONS FOR MULTI-PHASE DESIGNS.

12.1 Single-Phase. As was pointed out in Chapter 9, single-phase sampling is actually simple random sampling. Equations for sample size calculations are given in Chapter 2.

12.2 Two-Phase. The criterion for optimal sample size taken here will be to choose n and sampling fractions u_i; i = 1, 2, ..., I so that cost will be minimized for fixed variance or so that variance will be minimized for fixed cost. The procedure outlined is that presented by Cochran (1976, p. 331).

Assume a linear cost function of the form

$$C = C_1 n + \sum_{i=1}^{I} C_{2i} n_i \qquad (12.2.1)$$

where C = total cost,

C_1 = fixed cost per first-phase unit, and

C_{2i} = cost per second-phase unit in the ith first-phase stratum.

The expected cost is given by

$$E(C) = C^* = C_1 n + n \sum_{i=1}^{I} C_{2i} u_i W_i, \qquad (12.2.2)$$

where $W_i = N_i/N$.

The variance of $\hat{\mu}$ given by equation (9.3.3) may be rewritten as

$$V(\hat{\mu}) + \frac{S^2}{N} = \frac{1}{n}\left[S^2 + \sum_{i=1}^{I} W_i S_i^2 \left(\frac{1}{u_i} - 1\right)\right] \qquad (12.2.3)$$

where S^2 is the population variance, and

S_i^2 is the variance of the N_i units in stratum i.

Minimization of the product $C^*[V(\hat{\mu}) + \frac{S^2}{N}]$ insures that the optimal sample size criterion mentioned above is satisfied (Cochran, 1976, p. 281). This product does not involve n and it is possible to determine explicit expressions for the sampling fractions u_i that minimize $C^*[V(\hat{\mu}) + S^2/N]$. These expressions are

$$(12.2.4)\quad u_i = S_i \left[\frac{C_1}{C_{2i}\left(S^2 - \sum_{i=1}^{I} W_i S_i^2\right)} \right]^{\frac{1}{2}} \qquad \text{for } i = 1, 2, \ldots, I$$

The desired value for n may be obtained by substituting the u_i values obtained from Equation (12.2.4) into either Equation (12.2.2) or Equation (12.2.3). Since the values for W_i, S^2 and S_i^2 are unknown, it will be necessary to conduct a pilot survey to obtain estimates using

$$(12.2.5)\quad \frac{n_i}{n} \text{ to estimate } W_i,$$

$$(12.2.6)\quad s_i^2 = \sum_{j=1}^{m_i} (y_{ij} - \bar{y}_i)^2/(m_i - 1) \text{ to estimate } S_i^2$$

and

$$(12.2.7)\quad \frac{\hat{V}(\hat{\mu}) - \sum_{i=1}^{I} \frac{w_i}{n} s_i^2 (1/u_i - 1)}{\frac{1}{n} - \frac{1}{N}} \text{ to estimate } S^2$$

where $\hat{V}(\hat{\mu})$ is obtained from Equation (9.3.4).

12.3 Three-Phase. The objective of this section is to outline a procedure for the determination of sample size n, second-phase sampling fractions u_i and third-phase sampling fractions v_{ij} that satisfy the optimality criterion given in Section 12.2. The procedure is similar to that used in two-phase sampling except that it is possible to determine explicit sampling fraction values only in certain cases. Three cases will be treated in this section. The first assumes that the sampling fractions are equal and unspecified; the second assumes that the sampling fractions are unequal and specified; and the third assumes that the sampling fractions are unequal and unspecified.

As in two phase sampling, assume a linear cost function of the form

$$C = C_1 n + \sum_{i=1}^{I} C_{2i} m_i + \sum_{i=1}^{I} \sum_{j=1}^{J_i} C_{3ij} b_{ij} \tag{12.3.1}$$

C = total cost

C_1 = fixed cost per first phase unit,

C_{2i} = cost per second phase unit in the ith first phase stratum,

and C_{3ij} = cost per third phase unit in the jth second phase stratum in the ith first phase stratum.

The expected cost is given by

$$E(C) = C^* = n\left[C_1 + \sum_{i=1}^{I} C_{2i} \frac{Ni}{N} u_i + \sum_{i=1}^{I} \sum_{j=1}^{J_i} C_{3ij} \frac{N_{ij}}{N} u_i v_{ij}\right] \tag{12.3.2}$$

The variance of $\hat{\mu}$ is

$$V(\hat{\mu}) = \left(\frac{1}{n} - \frac{1}{N}\right) S^2 + \sum_{i=1}^{I} \frac{N_i}{N} \left(\frac{1}{u_i} - 1\right) \frac{S_i^2}{n} + \sum_{i=1}^{I} \left[\frac{1}{u_i} \sum_{j=1}^{J_i} \frac{N_{ij}}{N} \frac{S_{ij}^2}{n} \left(\frac{1}{v_{ij}} - 1\right)\right] \tag{12.3.3}$$

where S^2 is the population variance,

S_i^2 is the variance of the N_i units in the ith first-phase stratum, and

S_{ij}^2 is the variance of the N_{ij} units in the jth second phase stratum in the ith first phase stratum.

This equation may be rewritten as

$$V(\hat{\mu}) + \frac{1}{N} S^2 = \frac{1}{n}\left[S^2 + \sum_{i=1}^{I} \frac{N_i}{N} \left(\frac{1}{u_i} - 1\right) S_i^2 + \sum_{i=1}^{I} \frac{1}{u_i N} \sum_{j=1}^{J_i} N_{ij} S_{ij}^2 \left(\frac{1}{v_{ij}} - 1\right)\right] \tag{12.3.4}$$

As with two-phase sampling, it is desired to determine values for n and the sampling fractions that will minimize the product $C^*[V(\hat{\mu}) + S^2/N]$. In the first case to be considered, it is assumed that

and $\quad u_i = u$ for $i = 1, 2, \ldots, I$

$v_{ij} = v$ for $i = 1, 2, \ldots, I$; $j = 1, 2, \ldots, J_i$

The right hand sides of Equation (12.3.2) and Equation (12.3.3) can be simplified by adopting the following notation.

$$\text{Let } C_2 = \sum_{i=1}^{I} \frac{N_i}{N} C_{2i},$$

$$C_3 = \sum_{i=1}^{I} \sum_{j=1}^{J_i} \frac{N_{ij}}{N} C_{3ij},$$

$$S' = \sum_{i=1}^{I} \frac{N_i}{N} S_i^2,$$

$$S_3^* = S'' = \sum_{i=1}^{I} \sum_{j=1}^{J_i} \frac{N_{ij}}{N} S_{ij}^2,$$

$$S_1^* = S^2 - S'$$

and $\quad S_2^* = S' - S''.$

Then the problem is to minimize.

$$C^*[V(\hat{\mu}) + \frac{S^2}{N}] = [C_1 + C_2 u + C_3 uv][S_1^* + S_2^*/u + S_3^*/uv]. \tag{12.3.5}$$

If we assume that

$$S_1^* > 0 \text{ and } S_2^* > 0,$$

it can be shown (Sukhatme, p. 281) that choosing u to be,

$$u = \left[\frac{C_1 S_2^*}{C_2 S_1^*} \right]^{1/2}, \tag{12.3.6}$$

and v to be,

$$v = \left[\frac{C_2 \, S_3^*}{C_3 \, S_2^*} \right]^{-\frac{1}{2}}, \tag{12.3.7}$$

will minimize the product $C^*[V(\hat{\mu}) + \frac{S^2}{N}]$.

Note. To determine n, the values for u and v can be determined from Equation (12.3.6) and (12.3.7), and substituted into either Equation (12.3.2) or (12.3.4).

Note. It is possible for either of S_1^* and S_2^* to be negative. If one or the other is negative, then three-phase sampling is no more precise than two-phase sampling. If $S_2^* \leq 0$, and $S_1^* > 0$, then two-phase sampling should be used with the original first and second phases. If $S_1^* \leq 0$ and $S_2^* > 0$, then two-phase sampling should be used with the original first and third phases. If both are negative, then three-phase sampling is no more precise than one-phase (simple random) sampling.

The second case to be considered is the situation where the sampling fractions u_i and v_{ij} are specified. In this event, either Equation (12.3.2) or Equation (12.3.4) can be solved explicitly for n. For Equation (12.3.2),

$$n = \frac{C^*}{\left[C_1 + \sum_{i=1}^{I} C_{2i} \frac{N_i}{N} u_i + \sum_{i=1}^{I} \sum_{j=1}^{J_i} C_{3ij} \frac{N_{ij}}{N} u_i v_{ij} \right]}. \tag{12.3.8}$$

For Equation (12.3.4),

$$n = \frac{\left[S^2 + \sum_{i=1}^{I} \frac{N_i}{N} \left(\frac{1}{u_i} - 1 \right) S_i^2 + \sum_{i=1}^{I} \sum_{j=1}^{J_i} \frac{N_{ij}}{u_i N} S_{ij}^2 \left(\frac{1}{v_{ij}} - 1 \right) \right]}{(V^* + S^2/N)} \tag{12.3.9}$$

where V^* is the specified or desired value for $V(\hat{\mu})$.

The final case to be considered is the situation where the u_i and the v_{ij} are assumed to be unequal and unspecified. To solve this problem, it is necessary to resort to non-linear programming techniques. This problem may be stated as follows: find values of u_i and v_{ij} that minimize the objective function.

$$F = C^*[V(\hat{\mu}) + \frac{S^2}{N}] \tag{12.3.10}$$

subject to the constraints

$$0 < u_i \leq 1 \text{ for } i = 1, 2, \ldots, I$$

$$0 < v_{ij} \leq 1 \text{ for } j = 1, 2, \ldots, J_i;\ i = 1, 2, \ldots, I$$

Note that Equation (12.3.10) is equal to the product of the right hand sides of Equations (12.3.2) and Equation (12.3.4). This problem can be solved using a non-linear programming routine. To do this, the partial derivatives of Equation (12.3.10) with respect to the u_i and the v_{ij}, are required. An initial solution such as that given by Equation (12.3.6) and Equation (12.3.7) is also required. Once the non-linear programming solution is obtained, it can be substituted into either Equation (12.3.2) or Equation (12.3.4) to determine the sample size n.

As with two-phase sampling, the determination of n from either Equation (12.3.2) or Equation (12.3.4) requires that estimates of certain population parameters be obtained from data obtained in a pilot survey. These parameters can be estimated with the following formulae

$$\frac{n_i}{n} \qquad \text{to estimate } \frac{N_i}{N}, \tag{12.3.11}$$

$$\frac{m_{ij}}{m_i}\ \frac{n_i}{n} \qquad \text{to estimate } \frac{N_{ij}}{N}, \tag{12.3.12}$$

and

(12.3.13) $$\frac{\hat{V}(\hat{\mu}) - \sum_{i=1}^{I} \left(\frac{n_i}{n}\right)^2 \hat{V}_i}{\frac{1}{n} - \frac{1}{N}}$$ to estimate S^2,

where $\hat{V}_i$ is given by Equation (10.1.5) or Equation (10.1.7), and $\hat{V}(\hat{\mu})$ is given by Equation (10.1.3).

12.4 Four-Phase. As with two and three phase designs, the objective of this section is to outline a procedure for the determination of sample size n, second-phase sampling fractions u_i, third-phase sampling fractions v_{ij} and fourth-phase sampling fractions z_{ijk} that will satisfy the optimality criterion given in section (12.2). With the exception that the formulae required are different, this procedure is identical to that outlined in section (12.3) for three-phase sampling. The same three cases with respect to the sampling fractions that were considered in that section will be considered here.

As was done previously, assume a linear cost function of the form

(12.4.1) $$C = C_1 n + \sum_{i=1}^{I} C_{2i} m_i + \sum_{i=1}^{I} \sum_{j=1}^{J_i} C_{3ij} b_{ij} + \sum_{i=1}^{I} \sum_{j=1}^{J_i} \sum_{k=1}^{K_{ij}} C_{4ijk} c_{ijk}$$

where C = total cost,

C_1 = fixed cost per first-phase unit

C_{2i} = cost per second-phase unit in the ith first-phase stratum,

C_{3ij} = cost per third-phase unit in the jth second phase stratum in the ith first phase stratum, and

C_{4ijk} = cost per fourth-phase unit in the kth third-phase stratum in the jth second-phase stratum in the ith first-phase stratum.

The expected cost is given by

$$(12.4.2)\qquad E(C) = C^* = n\left[C_1 + \sum_{i=1}^{I} C_{2i}\frac{N_i}{N}u_i + \sum_{i=1}^{I}\sum_{j=1}^{J_i} C_{3ij}\frac{N_{ij}}{N}u_i v_{ij} + \sum_{i=1}^{I}\sum_{j=1}^{J_i}\sum_{k=1}^{K_{ij}} C_{4ijk}\frac{N_{ijk}}{N}u_i v_{ij} z_{ijk}\right].$$

The variance of $\hat{\mu}$ is

$$(12.4.3)\qquad V(\hat{\mu}) = \left(\frac{1}{n} - \frac{1}{N}\right)S^2 + \sum_{i=1}^{I}\frac{N_i}{N}\left(\frac{1}{u_i} - 1\right)\frac{S_i^2}{n} + \sum_{i=1}^{I}\frac{1}{u_i}\sum_{j=1}^{J_i}\frac{N_{ij}}{N}\left(\frac{1}{v_{ij}} - 1\right)\frac{S_{ij}^2}{n} + \sum_{i=1}^{I}\frac{1}{u_i}\sum_{j=1}^{J_i}\frac{1}{v_{ij}}\sum_{k=1}^{K_{ij}}\frac{N_{ijk}}{N}\left(\frac{1}{z_{ijk}} - 1\right)\frac{S_{ijk}^2}{n},$$

where S^2 is the population variance,

S_i^2 is the variance of the N_i units in the ith first-phase stratum,

S_{ij}^2 is the variance of the N_{ij} units in the jth second-phase stratum in the ith first-phase stratum, and

S_{ijk}^2 is the variance of N_{ijk} units in the kth third-phase stratum, in the jth second-phase stratum in the ith first-phase stratum.

It may be rewritten as

$$(12.4.4)\qquad V(\hat{\mu}) + \frac{S^2}{N} = \frac{1}{n}\left[S^2 + \sum_{i=1}^{I}\frac{N_i}{N}\left(\frac{1}{u_i} - 1\right)S_i^2 + \sum_{i=1}^{I}\sum_{j=1}^{J_i}\frac{1}{u_i}\frac{N_{ij}}{N}\left(\frac{1}{v_{ij}} - 1\right)S_{ij}^2 + \sum_{i=1}^{I}\sum_{j=1}^{J_i}\sum_{k=1}^{K_{ij}}\frac{N_{ijk}}{u_i v_{ij} N}\left(\frac{1}{z_{ijk}} - 1\right)S_{ijk}^2\right]$$

The first case, equality of the sampling fractions, is treated in the same manner as for three phase sampling. That is assume,

$$u_i = u \qquad \text{for } i = 1, \ldots, I$$

$$v_{ij} = v \qquad \text{for } i = 1, \ldots, I;\ j = 1, \ldots, J_i.$$

and $$z_{ijk} = z \qquad \text{for } i = 1, \ldots, I;\ j = 1, \ldots, J_i;\ k = 1, \ldots, K_{ij}.$$

Then Equation (12.4.3) may be rewritten as

$$\left[V(\hat{\mu}) + \frac{S^2}{N}\right] n = \left[S^2 - \sum_{i=1}^{I} \frac{N_i}{N} S_i^2\right] + \frac{1}{u}\left[\sum_{i=1}^{I} \frac{N_i}{N} S_i^2 - \sum_{i=1}^{I} \sum_{j=1}^{J_i} \frac{N_{ij}}{N} S_{ij}^2\right] + \frac{1}{uv}\left[\sum_{i=1}^{I} \sum_{j=1}^{J_i} \frac{N_{ij}}{N} S_{ij}^2 - \sum_{i=1}^{I} \sum_{j=1}^{J_i} \sum_{k=1}^{K_{ij}} \frac{N_{ijk}}{N} S_{ijk}^2\right] + \frac{1}{uvz}\left[\sum_{i=1}^{I} \sum_{j=1}^{J_i} \sum_{k=1}^{K_{ij}} \frac{N_{ijk}}{N} S_{ijk}^2\right] \tag{12.4.5}$$

which may be rewritten in simple notation as

$$[V(\hat{\mu}) + \frac{S^2}{N}]n = S_1^* + S_2^*/u + S_3^*/uv + S_4^*/uvz \tag{12.4.6}$$

In the same manner, Equation (12.4.2) may be rewritten as

$$\frac{C^*}{n} = C_1 + u \sum_{i=1}^{I} C_{2i} \frac{N_i}{N} + uv \sum_{i=1}^{I} \sum_{j=1}^{J_i} C_{3ij} \frac{N_{ij}}{N} + uvz \sum_{i=1}^{I} \sum_{j=1}^{J_i} \sum_{k=1}^{K_{ij}} C_{4ijk} \frac{N_{ijk}}{N} \tag{12.4.7}$$

which may be simplified to

$$\frac{C^*}{n} = C_1 + C_2 u + C_3 uv + C_4 uvz \tag{12.4.8}$$

If it is assumed that

$$S_1^* > 0,\ S_2^* > 0,\ S_3^* > 0$$

then it can be shown (Sukhatme, pg. 281) that the product

$$C^*[V(\hat{\mu}) + \frac{S^2}{N}] = [S_1^* + S_2^*/u + S_3^*/uv + S_4^*/uvz]\,[C_1 + C_2 u + C_3 uv + C_4\ uvz]$$

is minimized if u is chosen to be,

$$u = \left[\frac{C_1\ S_2^*}{C_2\ S_1^*}\right]^{\frac{1}{2}}, \tag{12.4.9}$$

and v is chosen to be

$$v = \left[\frac{C_2\ S_3^*}{C_3\ S_2^*}\right]^{\frac{1}{2}}, \tag{12.4.10}$$

and z is chosen to be,

$$z = \left[\frac{C_3\ S_4^*}{C_4\ S_3^*}\right]^{\frac{1}{2}}. \tag{12.4.11}$$

To determine n, substitute Equation (12.4.9), Equation (12.4.10), and Equation (12.4.11) into either Equation (12.4.5) or Equation (12.4.7).

Note As with three-phase sampling, one or more of the quantities S_1^*, S_2^*, and S_3^* may be negative. As before, this implies that four-phase sampling is no more efficient than three, two or one, phase sampling, depending on how many of the quantities are negative.

The second case, where the sampling fractions u_i, v_{ij} and z_{ijk} are all specified is handled in exactly the same way as in three-phase sampling. That is Equation (12.4.2) may be solved for n yielding.

(12.4.12)

$$n = \frac{C^*}{C_1 + \sum_{i=1}^{I} C_{2i} \frac{N_i}{N} u_i + \sum_{i=1}^{I} \sum_{j=1}^{J_i} C_{3ij} \frac{N_{ij}}{N} u_i v_{ij} + \sum_{i=1}^{I} \sum_{j=1}^{J_i} \sum_{k=1}^{K_{ij}} C_{4ijk} \frac{N_{ijk}}{N} u_i v_{ij} z_{ijk}}$$

The result for Equation (12.4.4) is

$$(12.4.13) \qquad n = \left[S^2 + \sum_{i=1}^{I} \frac{N_i}{N} \left(\frac{1}{u_i} - 1\right) S_i^2 + \sum_{i=1}^{I} \sum_{j=1}^{J_i} \frac{1}{u_i} \frac{N_{ij}}{N} \left(\frac{1}{v_{ij}} - 1\right) S_{ij}^2 + \sum_{i=1}^{I} \sum_{j=1}^{J_i} \sum_{k=1}^{K_{ij}} \frac{N_{ijk}}{u_i v_{ij} N} \left(\frac{1}{z_{ijk}} - 1\right) S_{ijk}^2 \right] / \left(V^* + \frac{S^2}{N}\right),$$

where V^* is the specified or desired value for $V(\hat{\mu})$.

The final case where the u_i, v_{ij} and z_{ijk} are all assumed to be unspecified is treated as a non-linear programming problem as in three-phase sampling. The problem may be stated as; find values of u_i, v_{ij} and z_{ijk} that minimize the objective funtion.

$$(12.4.14) \qquad F = C^*\left[V(\hat{\mu}) + \frac{S^2}{N}\right]$$

subject to the constraints

$0 < u_i \leq 1$ for $i = 1, \ldots, I$

$0 < v_{ij} \leq 1$ for $i = 1, \ldots, I$; $j = 1, \ldots, J_i$

$0 < z_{ijk} \leq 1$ for $i = 1, \ldots, I$; $j = 1, \ldots, J_i$; $k = 1, \ldots, K_{ij}$

Note that Equation (12.4.14) is equal to the product of the right hand sides of Equation (12.4.2) and Equation (12.4.5). To solve this problem, an initial solution such as that provided by Equation (12.4.9), Equation (12.4.10), and Equation (12.4.11) is required. Also, the partial derivatives of Equation (12.4.14) with respect to all of the sampling fractions must be computed. Once this problem is solved, the solution is subsituted into either Equation (12.4.2) or Equation (12.4.5) to determine n.

As with the designs discussed previously in this chapter, the determination of n from either Equation (12.3.2) or Equation (12.3.5) requires that estimates of certain population parameters be obtained from data obtained in a pilot survey. These parameters can be estimated with the following formulae,

(12.4.15) $\frac{n_i}{n}$ to estimate $\frac{N_i}{N}$,

(12.4.16) $\frac{m_{ij}}{m_i} \cdot \frac{n_i}{n}$ to estimate $\frac{N_{ij}}{N}$,

(12.4.17) $\frac{b_{ijk}}{b_{ij}} \cdot \frac{m_{ij}}{m_i} \cdot \frac{n_i}{n}$ to estimate $\frac{N_{ijk}}{N}$,

and

12.4.18) $$\frac{\hat{V}(\hat{\mu}) - \sum_{i=1}^{I} \left(\frac{n_i}{n}\right)^2 \hat{V}_i}{\left(\frac{1}{n} - \frac{1}{N}\right)}$$ to estimate S^2,

where $\hat{V}_i$ is given by Equation (11.1.4) and $\hat{V}(\hat{\mu})$ is given by Equation (11.1.3).

Appendix I

LEMMAS AND SYMBOLS

Lemma 1. Let X, Y be random variables and a be constant. Then,

(i) $E(X + Y) = E(X) + E(Y)$

(ii) $E(aX) = aE(X)$

(iii) $V(aX) = a^2V(X)$

Lemma 2. Under SRS the sample mean $\bar{y}$ is an unbiased estimate of the population mean $\bar{Y}$ (Cochran P.22)

Lemma 3. Under SRS $V(\bar{y}) = (\frac{1}{n} - \frac{1}{N})\ S^2$ (Cochran P. 23)

Lemma 4. Under SRS $\hat{V}(\bar{y}) = (\frac{1}{n} - \frac{1}{N})\ s^2$ (Cochran P. 26)

Lemma 5. Suppose a population consists of N primary sampling units (p.s.u.) and a SRS of size n is to be selected. Let $\hat{Y}_i$ be an unbiased estimator of the ith p.s.u. total based on any method of independent sampling. Also let $\hat{V}(\hat{Y}_i)$ be an unbiased estimator of the variance of $\hat{Y}_i$. Then

1) $\hat{Y} = \frac{N}{n} \sum_{i=1}^{n} \hat{Y}_i$ is an unbiased estimator of the population total Y.

2) $$\hat{V}(\hat{Y}) = \frac{N}{n}(N-n)\frac{\sum_{i=1}^{n}(\hat{Y}_i - \hat{\bar{Y}}_u)^2}{n-1} + \frac{N}{n}\sum_{i=1}^{n}\hat{V}(\hat{Y}_i)$$

is an unbiased estimate of $V(\hat{Y})$, where

$\hat{\bar{Y}}_u = \hat{Y}/N.$

Proof: Special case of a result by J.N.K. Rao. Unbiased variance estimation for multi-stage designs. Sankhyā C 37, 133-139, (1975).

Lemma 6. Suppose a population consists of N p.s.u.'s and a SRS of size n is to be selected. Let $\hat{Y}_i$ be an unbiased estimate of the ith p.s.u. total Y_i based on independent sampling at second and subsequent stages. Then,

(a) $\hat{Y} = \frac{N}{n} \sum_{i=1}^{n} \hat{Y}_i$ is an unbiased estimate of the population total Y.

(b) Bounds on an unbiased estimate of $V(\hat{Y})$ are given by:

$$\frac{N^2(N-n)}{Nn} s^2_{(\hat{Y}_i)}, \quad \frac{N^2}{n} s^2_{(\hat{Y}_i)}$$

where $$s^2_{(\hat{Y}_i)} = \frac{\sum_{i=1}^{n} (\hat{Y}_i - \hat{\bar{Y}}_.)}{n-1}$$

$$\hat{\bar{Y}}_. = \sum_{i=1}^{n} \hat{Y}_i/n$$

Rmk 1. The above theorem gives a method of getting 'bounds' on an unbiased estimate of variance, irrespective of how we sample in the second and subsequent stages as long as the subsamples are independent and we can get unbiased estimates of the ip.s.u. totals for i = 1, 2, ..., n. An example would be SRS in the 1st stage and systematic sampling in the second stage within each primary unit.

Rmk 2. The interval between the bounds tends to zero as $n/N \to 0$. Therefore, when $n/N << 1$ Lemma 6 gives an unbiased estimate of variance of the estimated total.

Lemma 7. Suppose a population consists of N p.s.u.'s and a probability proportional to P_i sample of size n with replacement is to be selected. Let $\hat{Y}_i$ be an unbiased estimate of the ith p.s.u. total based on independent sampling at the 2nd and subsequent stages. Then

(a) $\hat{Y} = \frac{1}{n} \sum_{i=1}^{n} \hat{Y}_i/P_i$ is an unbiased estimate of Y

(b) $\hat{V}(\hat{Y}) = \frac{s^2(\hat{Y}_i/P_i)}{n}$ is an unbiased estimate of the variance of $\hat{Y}$,

where $s^2(\hat{Y}_i/P_i) = \sum_{i=1}^{n} \frac{(Z_i - \bar{Z})^2}{n-1}$

where $Z_i = Y_i/P_i$ and $\bar{Z} = \sum_{i=1}^{n} Z_i/n$

Lemma 8. The sample mean of a single stage systematic sample is an unbiased estimator of the population mean. (Unbiased estimator of the variance of sample mean for a single stage systematic sample is not known).

Lemma 9. An estimate of the variance of the ratio (estimate)

$$\hat{R} = \frac{\sum_{n}^{n} M_i \bar{y}_i}{\sum_{n}^{n} M_i \bar{x}_i}$$

in two-stage designs is given by

$$\hat{V}(\hat{R}) \doteq \frac{1-f_1}{n\bar{x}^2} \sum^{n} \frac{M_i^2(\bar{y}_i - \hat{R}\bar{x}_i)^2}{n-1}$$

$$+ \frac{f_1}{n^2\bar{x}^2} \sum_{i=1}^{n} \frac{M_i^2(1-f_{2i})s^2_{d2i}}{m_i}$$

where $d_{ij} = y_{ij} - \hat{R}x_{ij}$

and $s^2_{d2i} = \frac{1}{n_i - 1} \sum^{m_i} (d_{ij} - \bar{d}_{i.})^2$

Proof: Cochran (2nd edition) Page 312.

On Multi Phase Sampling for Stratification

Consider a population of N sampling units u_{ij}, i = 1, 2, ...L and j = 1, 2, ..., N_i where prior to sampling the stratum identification (i) of the units and the stratum sizes (N_i) are unknown and Y_{ij} is a measurement of interest on unit U_{ij}. It is of interest

to estimate $\bar{Y} = \sum_{i=1}^{L} \sum_{j=1}^{N_i} Y_{ij}/N$ and $S^2 = \sum_{i=1}^{L} \sum_{j=1}^{N_i} (Y_{ij} - \bar{Y})^2/(N-1)$.

A first phase sample of size n is selected by simple random sampling without replacement. The selected units are classified into strata with the following results $u_{11}, u_{12}, \ldots, u_{1\,n_1}; u_{21}, u_{22}, \ldots, u_{2n_2}; \ldots; u_{L1}, \ldots, u_{Ln_L}$ where $n_1 + n_2 + \ldots + n_L = n$. The units are only classified into strata at phase one while actual measurement of the variable of interest is at a later phase of sampling. It is assumed that n is large enough so that $n_i \geq 2$. Consider the n units of the phase one sample as a stratified population with unknown "parameters" to be estimated. Define these "parameters" as $\bar{y}_i = \sum_{j=1}^{n} y_{ij}/n_i$ and

$s_i^2 = \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2/(n_i-1)$, i = 1, 2, ..., L where y_{ij} is the measurement value which would be obtained on unit u_{ij}. Also of interest are the "parameters"

$$\bar{y} = \sum_{i=1}^{L} \sum_{j=1}^{n_i} y_{ij}/n \text{ and } s^2 = \sum_{i=1}^{L} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2/(n-1).$$

In phase-two, sampling of the first phase "population" is performed. Any sampling procedure may be used in the ith stratum that is independent of the sampling in the other strata and for which estimators exist that are dependent on the ith

i) $\hat{\bar{y}}_i$ is an unbiased estimator of $\bar{y}_i$;

ii) $\hat{V}_i$ is an unbiased estimator of V_i, the variance of $\hat{\bar{y}}_i$ given the units selected in phase one; and

iii) $\hat{s}_1^2$ is an unbiased estimator of s_i^2.

Then unbiased estimators of $\bar{Y}$ and S^2 are given by (1) and (2) and an unbiased estimator of the variance of $\hat{\bar{Y}}$ is given by (3):

$$\hat{\bar{Y}} = \sum_{i-1}^{L} \frac{n_i}{n} \hat{\bar{y}}_i \quad , \tag{1}$$

$$\hat{S}^2 = \frac{\hat{V}(\hat{Y}) - \sum_{i=1}^{L} (\frac{n_i}{n})^2 \hat{V}_i}{(\frac{1}{n} - \frac{1}{N})} \quad , \tag{2}$$

$$\hat{V}(\hat{\bar{Y}}) = \frac{N-n}{Nn(n-1)} \left\{ \sum_{i=1}^{L} n_i (\hat{\bar{y}}_i - \hat{\bar{Y}})^2 + \sum_{i=1}^{L} (n_i - 1)\hat{s}_i^2 + \sum_{i=1}^{L} n_i \left[n_i (\frac{N-1}{N-n}) - 1 \right] \hat{V}_i \right\}. \tag{3}$$

The proofs of the unbiasedness of (1), (2) and (3) are given next. Conditional expectations will be used where $E(\hat{X}|I)$ means the expectation of $\hat{X}$ holding the sampling units selected in phase one fixed. E_I indicates expectation allowing the n sampling units selected at phase one to vary. Similarly V_I indicates variance as the phase one units vary. First

$$E(\hat{\bar{Y}}) = E_I \, E(\sum_{i=1}^{L} \frac{n_i}{n} \hat{\bar{y}}_i | I) = E_I \, (\bar{y}) = \bar{Y}$$

Next note that the variance of $\hat{\bar{Y}}$ can be written as

$$V(\hat{\bar{Y}}) = V_I\{E(\hat{\bar{Y}}|I)\} + E_I\{V(\hat{\bar{Y}}|I)\}$$

$$= (\frac{1}{n} - \frac{1}{N}) \, S^2 + E_I \left\{ \sum_{i=1}^{L} (\frac{n_i}{n})^2 \, V_i \right\}.$$

The second term in the above expression for $V(\hat{\bar{Y}})$ is the contribution to the variance of $\hat{\bar{Y}}$ from the second phase of sampling. Consider now the expected value of expression (3)

$$E(\hat{V}(\hat{\bar{Y}})) = \frac{N-n}{Nn(n-1)} \, E_I \left\{ E\left[\sum_{i=1}^{L} n_i (\hat{\bar{y}}_i - \hat{\bar{Y}})^2 | I \right] \right.$$

$$\left. + \sum_{i=1}^{L} (n_i - 1) s_i^2 + \sum_{i=1}^{L} n_i \left[n_i (\frac{N-1}{N-n}) - 1 \right] V_i \right\} .$$

Clearly

$$E \sum_{i=1}^{L} n_i(\hat{\bar{y}}_i - \hat{\bar{Y}})^2 | I = \sum n_i y_i^2 + \sum n_i V_i - n\, E(\hat{\bar{Y}}^2 | I)$$

$$= \sum n_i \bar{y}_i^2 + \sum n_i V_i - n(\bar{y}^2 + \sum (\frac{n_i}{n})^2 V_i)$$

Thus

$$E(\hat{V}(\hat{\bar{Y}})) = \frac{N-n}{Nn(n-1)} [E_I\{(n-1)s^2 + \frac{Nn(n-1)}{N-n} \sum_{i=1}^{L} (\frac{n_i}{n})^2 V_i\}] ,$$

since $(n-1)s^2 = (\sum_{i=1}^{L} n_i \bar{y}_i^2 - n\bar{y}^2) + \sum_{i=1}^{L} (n_i - 1)s_i^2$.

Now $E_I\ (s^2) = S^2$, hence

$$E(\hat{V}(\hat{\bar{Y}})) = (\frac{1}{n} - \frac{1}{N})S^2 + E_I \sum_{i=1}^{L} (\frac{n_i}{n})^2 V_i$$

as desired. It is noteworthy in the above argument that

$$E\ \{[\sum_{i=1}^{L} n_i(\hat{\bar{y}}_i - \hat{\bar{Y}})^2 + \sum_{i=1}^{L} (n_i - 1)\hat{s}_i^2 - \sum_{i=1}^{L} \frac{n_i}{n} (n - n_i)\hat{V}_i]|I\}$$

$$= (n-1)s^2.$$

Thus one has immediately that an unbiased estimator of S^2 is given by

$$\hat{S}^2 = \frac{1}{(n-1)} \left[\sum_{i=1}^{L} n_i(\hat{\bar{y}}_i - \hat{\bar{Y}})^2 + \sum_{i=1}^{L} (n_i - 1)s_i^2 - \sum_{i=1}^{L} \frac{n_i}{n} (n - n_i)\hat{V}_i \right].$$

$\hat{S}^2$ can be rewritten as

$$\hat{S}^2 = \frac{\hat{V}(\hat{\bar{Y}}) - \sum_{i=1}^{L} (\frac{n_i}{n})^2 \hat{V}_i}{(\frac{1}{n} - \frac{1}{N})}$$

LITERATURE CITED

Bickford, C. A., C. E. Mayer, and K. D. Ware. 1963. An efficient sampling design for forest inventory: the Northeastern forest resurvey. J. Forest. 61(11): 826-833.

Cochran, W. G. 1977. Sampling techniques. John Wiley and Sons, 3rd. ed. 428 pp.

Mendenhall, W., L. Ott, and R. L. Scheaffer. 1971. Elementary survey sampling. Duxbury Press, A Division of Wadsworth Publishing Co., Inc. 247 pp.

Rao, J. N. K. 1975. Unbiased variance estimation for multi-stage designs. Sankhyā: The Indian J. Statistics. 37(3): 133-139.

Sukhatme, P. V. 1957. Sampling theory of surveys with applications. Iowa State College Press, Ames, Iowa. 491 pp.